Duc Hoang Nguyen

The Role of Reinforcement Learning Control for Optimizing Building Energy Management Systems

School of Technology and Innovations
Master's thesis in
Smart Energy Programme

Vaasa 2024

UNIVERSITY OF VAASA
School of Technology and Innovations
Author: Duc Hoang Nguyen
Title of the Thesis: The Role of Reinforcement Learning Control for Optimizing Building Energy Management Systems
Degree: Master of Science in Technology
Programme: Smart Energy
Supervisor: Mazaher Karimi
Evaluator: Kimmo Kauhaniemi
Year: 2024 **Sivumäärä:** 97

ABSTRACT:

Modern energy management systems are increasingly challenged by the complexity, uncertainty, and high-dimensional data generated by advanced power systems. These challenges have driven a growing interest in integrating intelligent techniques such as machine learning (ML) and deep learning (DL) into energy management to improve system adaptability and efficiency. Among these approaches, reinforcement learning (RL) has emerged as a promising solution due to its ability to handle dynamic, sequential decision-making processes under uncertainty. RL has demonstrated its potential not only in energy management but also in related areas such as demand response, operational control, and renewable energy integration.

This research delves into the development of RL-based frameworks for intelligent energy management in grid-interactive buildings, with a special focus on the integration of electric vehicles (EVs) as distributed energy resources (DERs). Firstly, the thesis carries out a systematic review of the foundational principles of RL and its diverse applications within the domain of power systems.

Subsequently, a data-driven framework leveraging the Soft Actor-Critic RL algorithm is proposed to enable prosumers to reduce energy costs, enhance grid stability, improve renewable energy utilization, and maintain user comfort. Simulation results highlight the effectiveness of the proposed framework, showing significant performance gains over state-of-the-art control strategies in terms of cost efficiency, CO_2 emissions reduction, and grid resilience. Additionally, the study provides a critical evaluation of the practical challenges and opportunities of implementing RL-based systems in real-world scenarios.

The insights gained highlight the transformative potential of RL in enabling adaptive and sustainable energy management practices. By addressing both the technical complexities and real-world applications, this research advances the understanding of intelligent energy systems and underscores the importance of RL in meeting the growing demands of modern energy infrastructures while promoting sustainability and economic viability.

KEYWORDS: Artificial Intelligence, Deep Learning, Energy Management Systems, Reinforcement Learning, Renewable Energy.

Contents

FIGURES

TABLES

Abbreviations

AI: Artificial Intelligence
ANN: Artificial Neural Network
BEM: Building Energy Management
BESS: Battery Energy Storage System
DER: Distributed Energy Resource
DNN: Deep Neural Network
DPG: Deterministic Policy Gradient
EMS: Energy Management System
EV: Electric Vehicle
EVC: Electric Vehicle Charger
G2V: Grid To Vehicle
GEB: Grid-Interactive Efficient Building
HEM: Home Energy Management
HVAC: Heating, Ventilation, Air-Conditioning
iEMS: Intelligent Energy Management System
IoT: Internet of Things
KPI: Key Performance Indicator
MDP: Markov Decision Process
MDS: Multidimensional Scaling
MPC: Model Predictive Control
PV: Photovoltaic
RBC: Rule-Based Control
RES: Renewable Energy Sources
RL: Reinforcement Learning
SAC: Soft Actor-Critic
V2G: Vehicle To Grid

1 Introduction

After decades of development, power systems have become highly complex networks of electronics and electrical devices. These days, economic, technological, environmental, and political motivations have driven the transformation of conventional grids into more sophisticated, reliable, and sustainable grids (Ipakchi and Albuyeh, 2009). Within that landscape of modern energy systems, microgrids have emerged as a transformative solution, offering localized and resilient power generation and distribution capabilities. Microgrids function on a smaller scale than conventional centralized power grids, providing power to certain communities, regions, or even individual buildings (Parhizi et al., 2015). Microgrids are revolutionizing the production, distribution, and use of power by integrating various renewable energy sources, energy storage devices, and cutting-edge control technology.

Although the deployment of microgrids is significantly growing, the high penetration of renewables and integration of distributed energy resources (DERs) including electric vehicles (EVs), as well as the bi-directional energy flow and information, have brought many challenges (Vineetha and Babu, 2014). First, the large penetration of intermittent resources brings operational challenges and impacts on power quality and system stability due to the uncertain nature of renewable resources. For instance, a wind turbine generates fast ramping as wind speed changes while also creating sudden power cutoff at high wind speeds to protect its blades and turbine. Secondly, the explosion of information and its growing complexity causes problems for grid operators in analyzing and making rational decisions regarding grid operation (Hossain et al., 2019).

To tackle these challenges, many studies have been conducted in different aspects such as architecture, power electronics, security, demand response, energy scheduling strategy, and energy management systems. Research on effective energy management strategies has been a primary focus and a prevailing topic of interest in the field (Nakabi and Toivanen, 2021).

The current energy management system (EMS) must be extensively re-engineered into an integrated and intelligent eco-system that handles not only microgrids' components but also utilizes data from different energy sources, hardware power, and even the electricity spot market price. Therefore, modelling techniques and associated essential enabling technologies remain hot issues that demand extensive scientific study. Especially, there is a growing need for more sophisticated and adaptive approaches to artificial intelligence (AI) – based EMS to maximize energy flexibility while minimizing the operation cost.

This thesis aims to design and simulate an intelligent energy management system on the prosumer side that integrates the approach of energy consumption and generation predictions with deep learning and reinforcement learning (RL) techniques to select the optimal actions and achieve the best performance by evaluating multiple key performance indicators (KPIs). The research system must be flexible to handle such complex energy systems including renewable energy sources (solar), batteries, and electric vehicle charging stations.

We structure this work as follows: In Chapter 2, we provide the study background by exploring global energy challenges, discussing advanced energy management approaches, and emphasizing RL's relevance in energy optimization. Chapter 3 outlines the research questions, methodology and research process. Chapter 4 details the literature review presenting key findings on energy management strategies and RL applications, and identifying gaps that the study aims to address. Based on the literature review, in Chapter 5, we propose an RL-based iEMS for the simulation study including system modelling, the design of an RL control algorithm and performance evaluating methods. Chapter 6 describes the simulation work evaluating the performance of the optimized algorithm. We then reflect on the application of RL in energy management and address challenges in implementing RL-based solutions in real-world scenarios in Chapter 7.

2 Study Background

2.1 Global Energy Challenges and Sustainability

Buildings constitute a significant source of greenhouse gas emissions, with their carbon footprint steadily increasing. Globally, buildings account for over 34% of the total energy demand and approximately 37% of CO2 emissions (United Nations, 2022). The same report also found that:

- CO_2 emissions witnessed a 2% increase from pre-pandemic levels.
- Operational energy demand for heating, cooling, lightning, and other appliances grew by 3% from 2019.
- Notable uptick of 16% in investments to enhance energy efficiency in 2021. However, this advancement was counteracted by an unparalleled expansion in the floor space, thus negating the positive impacts of the energy efficiency measures.

Mitigating the carbon footprint of the building sector is now more crucial than ever in reaching the climate objectives established by multiple nations in the Paris Agreement, aimed at addressing climate change.

Yet sustainability goals, which are spearheaded by global institutions including the United Nations (UN) and the Paris Agreement, are meant to tackle the urgent issues brought about by climate change. The urgent necessity of shifting to clean, renewable energy sources and improving energy efficiency in order to mitigate the effects of climate change is emphasized by the UN Sustainable Development Goals (SDGs), specifically Goal 7 (Affordable and Clean Energy) and Goal 13 (Climate Action). The worldwide agreement on lowering greenhouse gas (GHG) emissions, encouraging energy conservation, and guaranteeing sustainable economic growth is reflected in these objectives.

Climate change concerns center around the dramatic rise in global temperatures, which is largely attributed to the burning of fossil fuels and deforestation. The resulting increase in carbon dioxide (CO_2) and other GHGs in the atmosphere has led to more

frequent and severe extreme weather events, such as hurricanes, wildfires, and floods. These environmental changes threaten ecosystems, biodiversity, and human health, while also causing significant economic damage.

Figure 1. The UN Sustainable Development Goals (United Nations, 2015).

Improving energy efficiency in buildings, industries, and transportation is crucial for reducing global energy consumption and CO_2 emissions. Sustainable technologies and intelligent energy management systems, especially those powered by AI and machine learning, are essential for meeting climate goals and reducing energy waste. As smart buildings become more prevalent, integrating technologies like AI, IoT, and building automation, researchers are developing advanced control algorithms to optimize energy use and maintain comfort. These systems must manage not only traditional loads like HVAC, but also incorporate renewable energy, energy storage, and electric vehicle charging.

2.2 Reinforcement Learning

2.2.1 Overview of Reinforcement Learning

Reinforcement learning (RL) is a field of machine learning in which an AI system, often called an agent, learns by engaging with its environment E and getting feedback based on the given actions. This feedback comes in the form of rewards (R), intending to maximize the cumulative reward (Sutton and Barto, 2018). The process revolves around two main components: a policy (π_t) that guides the agent's actions (a_t) based on the state (s_t), and a learning algorithm that updates the policy to improve the cumulative reward over time.

RL aims to train the agent to accomplish a task in an unfamiliar environment. The agent interacts by receiving observations and rewards from the environment and responding with actions. The reward emphasizes how effective the action is in achieving the task's objective.

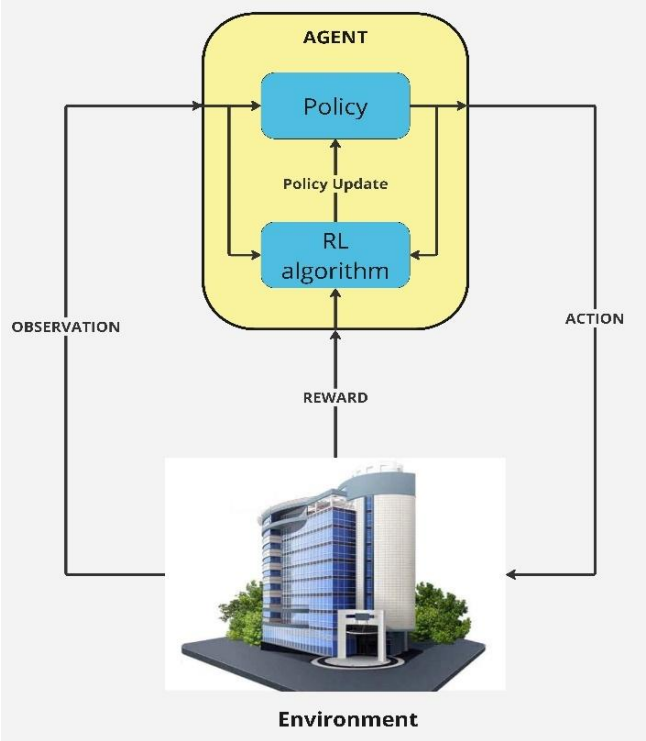

Figure 2. Illustration of a reinforcement learning model.

In a broader context, we know that machine learning methods are categorized into supervised and unsupervised learning. In supervised learning, both input data and corresponding output labels are provided, with regression predicting continuous values and classification predicting discrete categories (Jung, 2022). The model learns by minimizing a loss function to reduce the error between predicted and actual labels. In unsupervised learning, models detect patterns in data without predefined labels, useful for exploratory analysis or generating class labels for later supervised learning (Jo, 2021).

Reinforcement learning (RL) shares elements with both, as it involves learning from observations, but instead of labeled data, an agent interacts with an environment and is rewarded or penalized based on actions (Sutton and Barto, 2018).

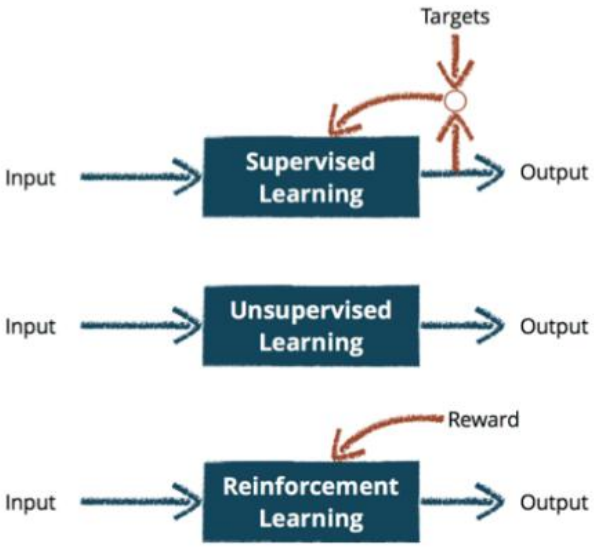

Figure 3. Machine learning techniques (Nagy et al., 2018).

A unique challenge in RL, unlike other types of learning, is managing the balance between exploration and exploitation. To maximize rewards, an agent must favour actions it has previously found to be effective. However, to identify these actions, the agent also needs to try new, unexplored territories (Sutton and Barto, 2018). This creates tension: the agent must exploit its current knowledge to gain rewards but also explore new possibilities to enrich the knowledge base and to improve future action choices. Focusing solely on exploration or exploitation would lead to failure, so the agent must experiment

with different actions while increasingly prioritizing those that seem most effective. In stochastic environments, each action needs to be tested multiple times to accurately predict the expected reward. This exploration-exploitation dilemma has been studied in great detail by mathematicians over the years, but it is worth noting that this issue does not arise in the setting of supervised learning as it is typically defined. Another challenge is the curse of dimensionality, where the number of potential states and actions in a complex environment becomes excessively large, making it challenging to determine an optimal policy (Nagy et al., 2018).

2.2.2 Reinforcement Learning Applications in General

RL has gained significant attention for its applications in real-world scenarios across various industries (Sutton and Barto, 2018). Numerous papers have suggested the use of Deep RL for autonomous driving (Kiran et al., 2022). In self-driving cars, multiple factors must be examined, including speed limits in different areas, identifying drivable zones, and avoiding collisions. RL can be applied to various autonomous driving tasks, such as route optimization, motion strategy, dynamic pathfinding, controller tuning, and developing scenario-based principles for highway driving (Kiran et al., 2022). For instance, RL can be used to learn automatic parking strategies, while lane changes can be handled using Q-Learning. Overtaking can be achieved by training an overtaking policy that avoids collisions and maintains a consistent speed afterwards. AWS DeepRacer, a self-driven racing vehicle created to test RL on a real track, is a fascinating illustration of RL in action (“AWS DeepRacer,” 2024).

In industry, robots powered by RL are employed to perform various tasks. These robots not only operate more efficiently than humans but are also capable of handling dangerous tasks that would pose risks to people. A prime example is DeepMind’s deployment of AI agents to regulate the cooling systems in Google Data Centers, which led to a 40% decrease in energy costs. The AI system now manages the centers autonomously, without requiring human involvement (“Google DeepMind,” 2024).

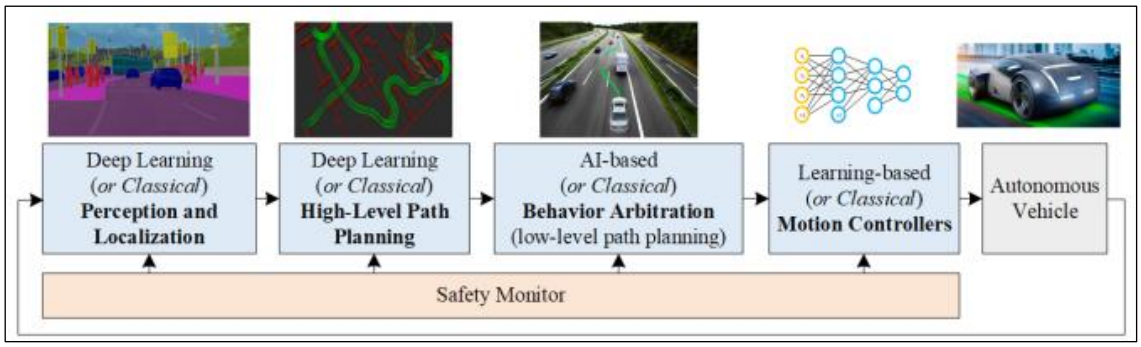

Figure 4. A modular perception-planning-action pipeline for deep learning-based self-driving cars (Grigorescu et al., 2020).

Another impressive application of RL is in gaming. Google DeepMind team used RL to develop artificial intelligence capable of playing complex games such as chess, Go, and shogi (Japanese chess). This approach was instrumental in creating AlphaGo, the first AI to defeat a professional human Go player. Building on that success, DeepMind developed AlphaZero, a deep neural network agent that taught itself to play chess at a level advanced enough to surpass the Stockfish chess engine in just four hours (Silver et al., 2017).

AlphaZero operates with only two key components: a neural network and an algorithm known as Monte Carlo Tree Search. This contrasts with the brute-force computing approach of Deep Blue, which, in 1997, when it beat world chess champion Garry Kasparov, processed 200 million possible chess positions per second. However, unlike Deep Blue, the workings of deep neural networks like AlphaZero are less transparent, limiting our understanding of their decision-making processes (Silver et al., 2017).

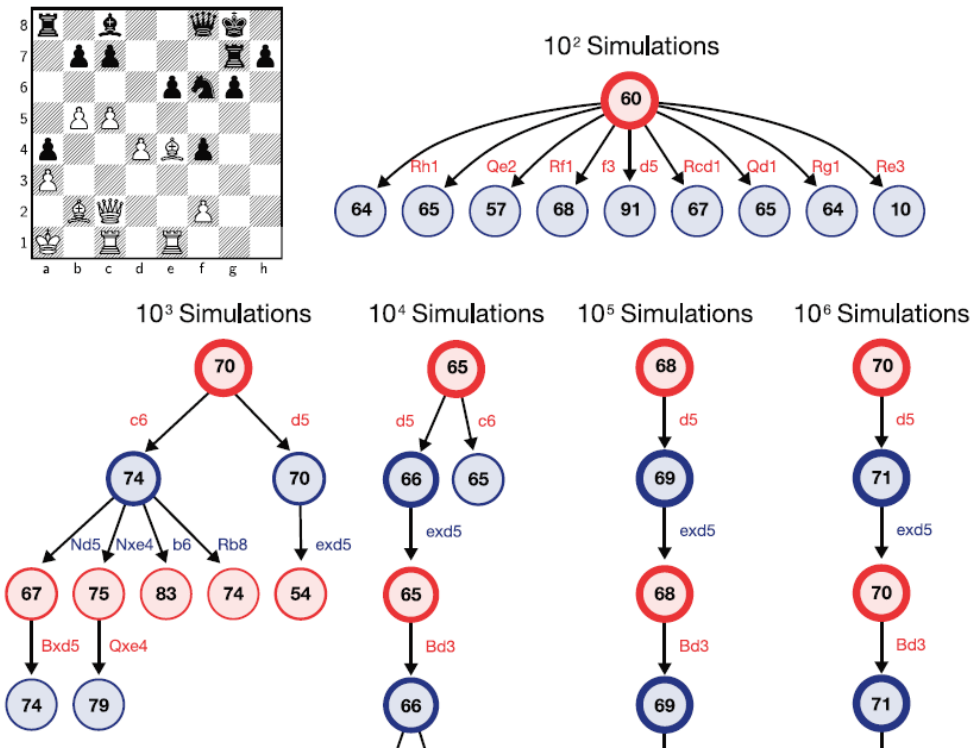

Figure 5. RL-based AlphaZero in chess games (Silver et al., 2017).

3 Research Objectives and Process

3.1 Research Objectives and Questions

Based on the study background in section 2, the goal of this study is to address the following research questions:

Q1: What could be the main research trends highlighting the current state of RL algorithms in the context of building energy management?

Q2: How can an intelligent energy management system (iEMS) be designed and developed for prosumers, integrating renewable energy and EVs using RL techniques?

Q3: How does the performance of the proposed iEMS compare to existing state-of-the-art control methods?

Q4: What are the key recommendations for the effective implementation and deployment of iEMS in real-world prosumer-oriented environments?

3.2 Research Process

The focus of this research is to explore the application of RL in the field of building energy management and its integration with prosumers (entities that both produce and consume energy). With increasing complexity in energy systems due to the inclusion of RES, such as solar PV, and EVs, traditional energy management methods struggle to optimize decisions in real time. An intelligent Energy Management System (iEMS), driven by RL techniques, can address this challenge by continuously learning and adapting to dynamic energy environments.

Details of the research process are outlined below.

3.2.1 Literature Review and Trend Identification (Step 1)

Objective (Q1): Understand the current state and trends in RL algorithms for building energy management. Conduct a comprehensive literature review of RL applications in

energy management, identifying key algorithms, such as Q-Learning, Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), and Actor-Critic methods. Analyze recent studies that address different aspects of energy optimization, including load shifting, demand response, and energy storage. Highlight trends and research groups, such as the use of multi-agent systems, deep RL for continuous control, and model-free RL in handling real-time, stochastic energy environments.

Deliverable: A systematic review that maps the evolution of RL algorithms in building energy systems, identifying strengths, limitations, and gaps in existing research.

3.2.2 Design and Development of an iEMS (Step 2)

Objective (Q2):

Develop an iEMS for prosumers that integrates RES and EVs using RL techniques.

- **System Design:**

  Define the architecture of the iEMS, incorporating key components such as solar PV, battery storage, EV charging stations, and building loads.

- **RL Algorithm Selection:**

  Based on findings from Step 1, select and customize RL algorithms for the specific energy management tasks. For instance, use Deep RL (DRL) to manage decision-making in continuous action spaces, such as energy storage and load shifting.

  Model prosumers' energy production (from solar PV) and consumption, and EV charging/discharging schedules.

- **Simulation Environment:**

  Use platforms like CityLearn to simulate urban environments and building interactions, allowing for the testing of RL-based iEMS in scenarios with fluctuating renewable energy, variable EV charging demands, and real-time energy prices.

Deliverable:

A functional iEMS prototype that leverages RL algorithms to dynamically optimize energy flows between renewable sources, EVs, storage systems, and the grid.

3.2.3 Performance Evaluation and Benchmarking (Step 3)

Objective (Q3):

Evaluate the performance of the proposed iEMS compared to existing control methods.

- **Performance KPIs:**
  Define main KPIs for comparison: net energy cost, emissions, average daily peak, (1 - load factor), ramping, and EV user comfort.
- **Benchmarking:**
  Compare the RL-based iEMS with state-of-the-art control methods, such as baseline, rule-based, and default RL control.
- **Experiments:**
  Run simulations to test the iEMS's adaptability and robustness.

Deliverable:

A performance evaluation report detailing how the RL-based iEMS outperforms (or performs comparably to) existing methods in terms of efficiency, scalability, and real-time adaptability.

3.2.4 Conclusion and Future Direction (Step 4)

Objective (Q4):

Identify and suggest best practices and challenges for deploying iEMS in real-world prosumer environments.

Deliverable:

A set of guidelines and recommendations for the effective implementation and deployment of RL-based iEMS in real-world prosumer settings, addressing technological, regulatory, and user-related challenges.

4 Literature Review

4.1 Literature Extraction Strategy and Data Set

The literature review session proceeds through a series of steps and employs a systematic review technique for reviewing published papers, with the ultimate goal of identifying key concepts, terminology, and research gaps in existing knowledge. The study eliminates the possibility of bias by using a clear, strict, and reliable technique to provide objective and repeatable results. The review's primary data source is derived from the Web of Science's (WOS) most current publications. The following are the extraction criteria:

1) Publications' titles or abstracts contain the keywords "building" or "home" or "house" and "energy management" and "reinforcement learning".
2) Articles published during the period between 2013 to 2024 (the most ten years).
3) Articles devoted only to examining the application of RL algorithms in building, household, or home energy management systems.

The data extraction process provides a data set of 143 articles that are published by different publishers such as Elsevier (69), IEEE (40), Mdpi (24), and others (10).
The criteria and data extraction process are demonstrated in Figure 6 below.

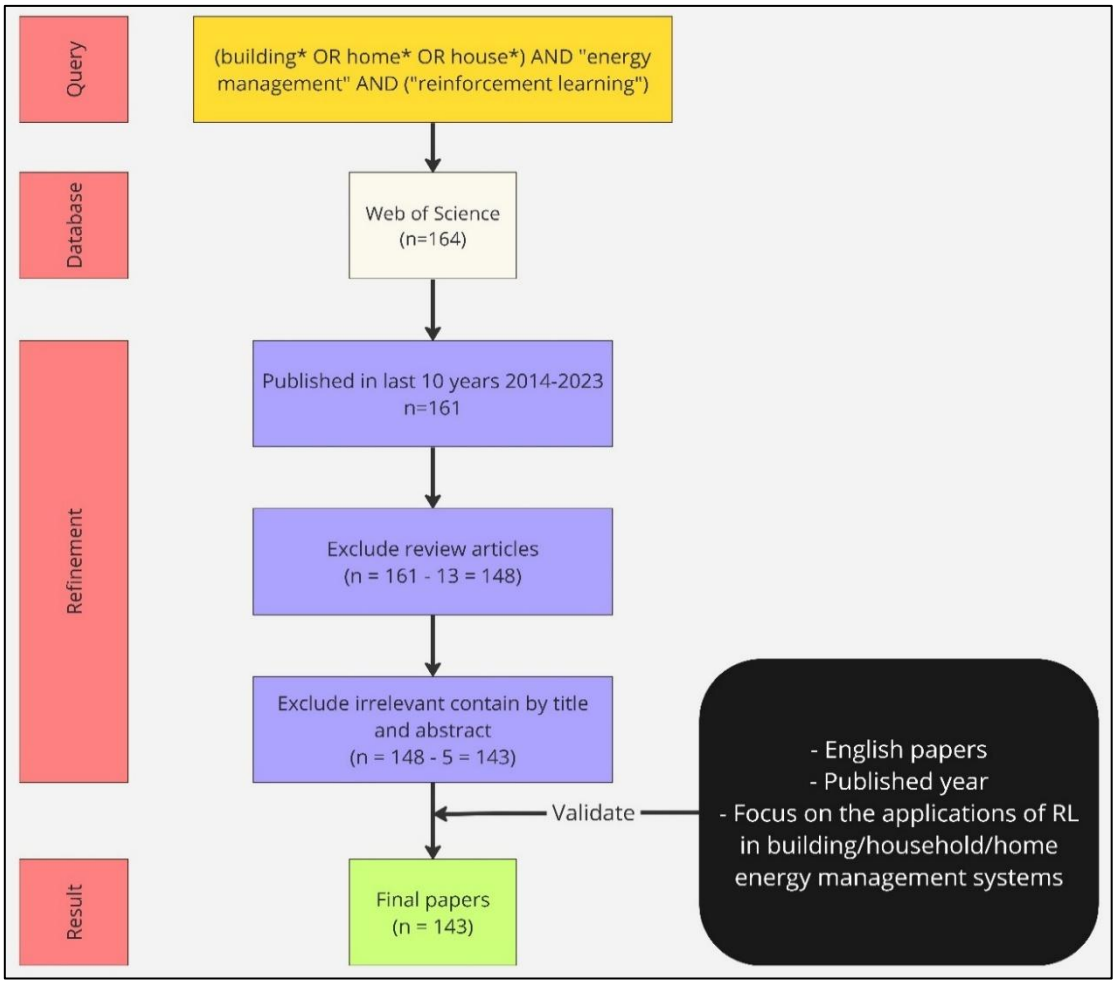

Figure 6. Literature extraction process.

The visual representation of keyword networks (Figure 7) emphasizes key terms such as "reinforcement learning," "demand response," "home energy management," "deep reinforcement learning," "energy management," and "optimization." Additionally, categorizing these keywords (as depicted in Figure 8) reveals a predominant focus in the literature on applying RL and artificial intelligence to optimize energy system components (including HVAC, EVs, RES, batteries, energy storage, and home appliances) within smart buildings or homes. These optimizations aim to achieve various objectives related to energy costs, consumption, and user comfort. Subsequent sections will delve into the review methodology, process, and findings in greater detail.

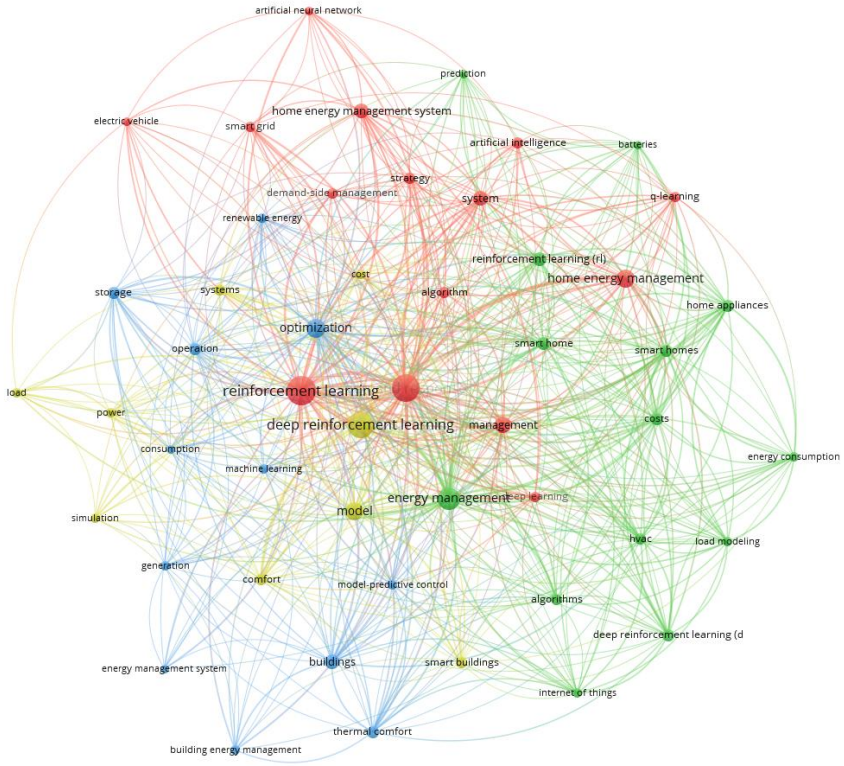

Figure 7. Article Network Visualization using Keyword Data.

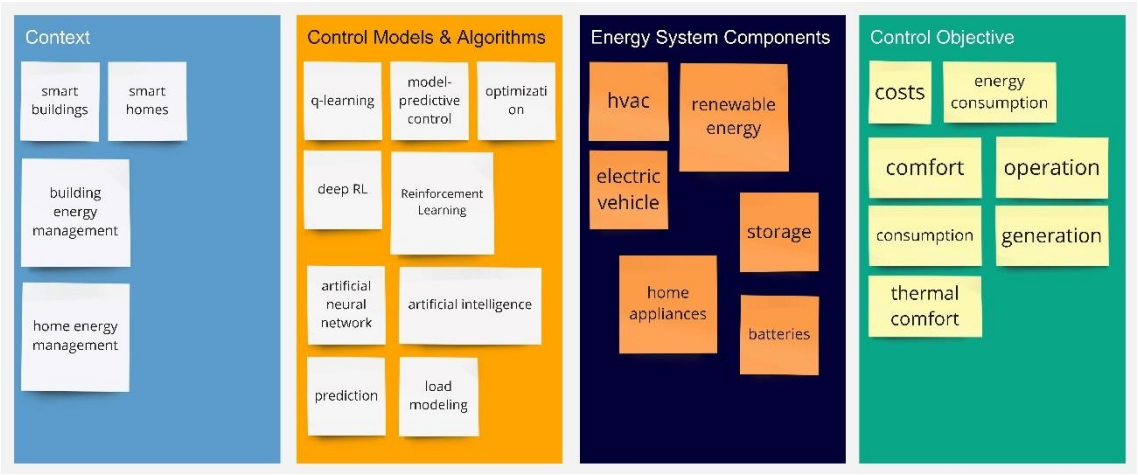

Figure 8. Classification of keywords.

4.2 Review Methodology

This review employs co-citation analysis to perform thematic analysis and identify primary research themes and emerging trends within those themes.

Small (1973) introduced co-citation analysis as a viable approach for identifying subject similarity and analyzing the evolutionary trends within the research field. Two documents (earlier) are co-cited when referenced in the bibliography of a third document (later). If papers A and B are both referenced by paper C, they are considered related, even if they do not directly reference each other. The strength of the co-citation increases if the earlier papers (A and B) are referenced by numerous other papers.

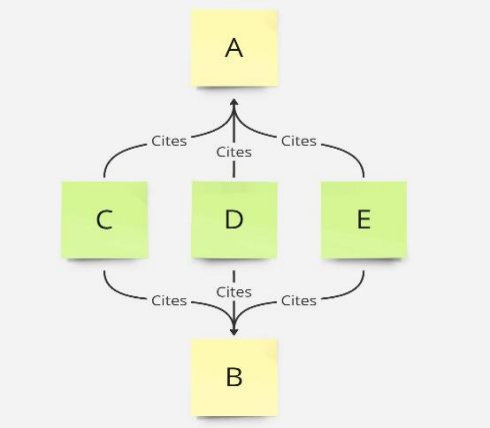

Figure 9. An illustration of co-citation coupling

Co-citation analysis stands out as a distinctive approach to exploring the cognitive framework of scientific fields (Small, 1973). Researchers can utilize the method to establish the connection between two or more studies by analyzing the overlap in the sources that cite them. Essentially, the more shared references between studies, the greater the similarity of their knowledge foundation (Schneider and Borlund, 2004). As multiple authors co-cite the same papers, clusters of research emerge. Within these clusters, the co-cited papers typically revolve around common themes. Hence, the findings of the analysis assist in identifying research streams and their future evolution. Such an approach is ideal for this review since it aims to investigate the present status, research gaps, and prospects of RL applications in the context of buildings and home energy management. Microsoft Excel and IBM SPSS are utilized as the tools for conducting the review.

4.3 Data Analysis Procedure

In this section, we explain a detailed process to carry out the co-citation analysis for the review work. Figure 10 outlines the data analysis procedure step-by-step.

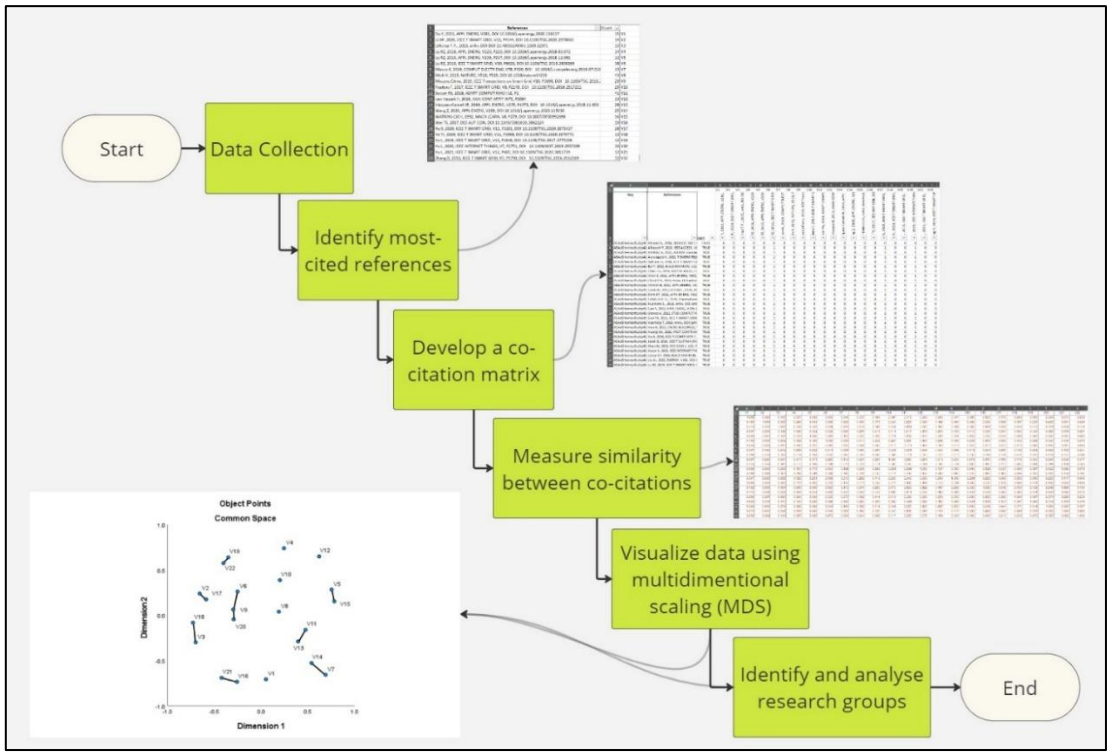

Figure 10. Co-citation analysis process.

The initial step is the data collection which has been explained above in the Literature Extraction Strategy and Data. This resulted in a set of 143 papers in the WOS database, containing 6,403 citations. Then we identified the most frequently cited references using citation frequency, a standard approach in studies concerning knowledge structure (Chabowski et al., 2013).

The citation frequencies highlight the most influential work, which holds a significant impact on the extracted papers. The subsequent step involves creating a co-citation matrix that identifies the extracted articles based on their shared citations. Technically speaking, the outcome is a 2-dimensional binary matrix mapping between the original papers and those most cited references. Next, we measure the similarity (distances)

between the references using the Ochiai coefficient since our data is in binary format (Zhou and Leydesdorff, 2016). The following step is to visualize and cluster highly similar articles together using multidimensional scaling (MDS), which forms a two-dimensional spatial layout representing the knowledge structure of the extracted articles (B.Kruskal and Wish, 1978). Finally, research groups or clusters are established by linking pairs of articles that are within a distance threshold of 0.25 or lower, ensuring the interpretability and significance of the results (Tsai and Wu, 2010).

4.4 Results

Based on the citation data, the most extensively cited references have been identified and shown in Table 1. It is important to highlight that the references listed here are not selected based on the highest number of citations from the WOS database. Instead, they are chosen based on their prominence within the 143 publications gathered for this review. Moreover, the overview indicates that the book "Reinforcement Learning: An Introduction" authored by Sutton and Barto, and published by MIT Press, serves as a cornerstone in the establishment of RL techniques, being the most frequently cited publication in the field.

We assign variables (V1 to V22) to each of the most-cited publications, enabling us to visualize their distances in a two-dimensional space. This facilitates the identification of research groups comprising publications that are closely related and share common knowledge.

Table 1. Most cited publications by the extracted data set

Code	References	Type	Source	Citation	Focus Overview
V1	(Du et al., 2021)	Research Paper	Applied Energy	169	An optimal control approach with the DDPG algorithm for a residential HVAC system across multiple zones seeks to reduce energy expenses while ensuring occupants' comfort.
V2	(Li et al., 2020)	Research Paper	IEEE Transactions on Smart Grid	105	A deep RL algorithm for demand response strategy for optimizing home appliance scheduling, considering uncertainties in resident behaviour, real-time electricity prices, and outdoor temperature.
V3	(Lillicrap et al., 2019)	Research Paper	arXiv	15176	Advanced algorithms for continuous state and action spaces, extending the Deep Q-Learning technique.
V4	(Lu et al., 2018)	Research Paper	Applied Energy	377	Electricity Service Providers' decision-making model based on Q-Learning to achieve Service Providers' and customers' benefit while balancing demand response in the market.
V5	(Lu and Hong, 2019)	Research Paper	Applied Energy	261	A real-time incentive-based demand response technique for smart grid systems utilizing RL and DNN.
V6	(Lu et al., 2019)	Research Paper	IEEE Transactions on Smart Grid	229	A multi-agent RL and ANN technique to ensure energy demand response with optimal control of different home appliances.
V7	(Mason and Grijalva, 2019)	ReviewPaper	Computers & Electrical Engineering	155	The application of RL in developing autonomous building energy management systems.
V8	(Ruelens et al., 2017)	Research Paper	IEEE Transactions on Smart Grid	226	Applying batch RL techniques to address the demand response problem in residential electrical water heaters.
V9	(Mocanu et al., 2019)	Research Paper	IEEE Transactions on Smart Grid	562	Deep Q-learning and deterministic policy gradient methods to perform online optimization schedules for building energy management.
V10	(Mnih et al., 2015)	Research Paper	Nature	29030	Explanation regarding network structure, parameters, and algorithm of the Deep Q-Network (DQN) and demonstration applications on multiple human games.
V11	(Sutton and Barto, 2018)	Book	MIT press	69561	In-depth exploration of RL covering fundamental concepts, theoretical foundations, and practical implementations of RL techniques.
V12	(Van Hasselt et al., 2016)	Conference Proceeding	AAAI Conference on Artificial Intelligence	8583	A theoretical foundation and empirical results of Double DQN, extending the success of the DQN algorithm.
V13	(Vázquez-Canteli and Nagy, 2019)	Review Paper	Applied Energy	428	The role of RL in demand response applications within the smart grid context.
V14	(Wang and Hong, 2020a)	Review Paper	Applied Energy	243	A review of the applications of RL in building controls focusing on examining algorithms, state, action, reward, and environment.
V15	(Watkins and Dayan, 1992)	Research Paper	Machine Learning	7248	Theoretical basis of the model-free Q-Learning algorithm.
V16	(Wei et al., 2017)	Conference Proceeding	IEEE/IET Electronic Library (IEL)	33	DRL-based approaches for HVAC systems in buildings to reduce energy costs and ensure user comfort.
V17	(Xu et al., 2020)	Research Paper	IEEE Transactions on Smart Grid	212	A data-driven approach based on neural networks and RL to achieve demand response at the household level. Real-time electricity prices, PV generation, and EV are considered.
V18	(Ye et al., 2020)	Research Paper	IEEE Transactions on Smart Grid	116	A combination of the deep deterministic policy gradient (DDPG) algorithm with a prioritized experience replay strategy to optimize energy usage for residential multi-energy systems.
V19	(Yu et al., 2019)	Research Paper	IEEE Transactions on Smart Grid	103	An online energy management algorithm for minimizing energy cost and thermal discomfort cost within a smart home environment.
V20	(Yu et al., 2020)	Research Paper	IEEE Internet of Things Journal	225	A DDPG-based energy management algorithm to control HVAC systems and ESS within smart homes without a building thermal dynamics model.
V21	(Yu et al., 2021b)	Research Paper	IEEE Transactions on Smart Grid	147	A multi-agent deep RL algorithm to control HVAC in multi-zone commercial buildings without buildings' thermal dynamics models.
V22	(Zhang et al., 2016)	Research Paper	IEEE Transactions on Smart Grid	167	A learning-based demand response mechanism for home energy management.

Using MDS analysis via SPSS, the most cited works have been visualized in two-dimensional space as shown in Figure 11. More frequently co-citations present greater commonality in the knowledge foundation and closer proximity.

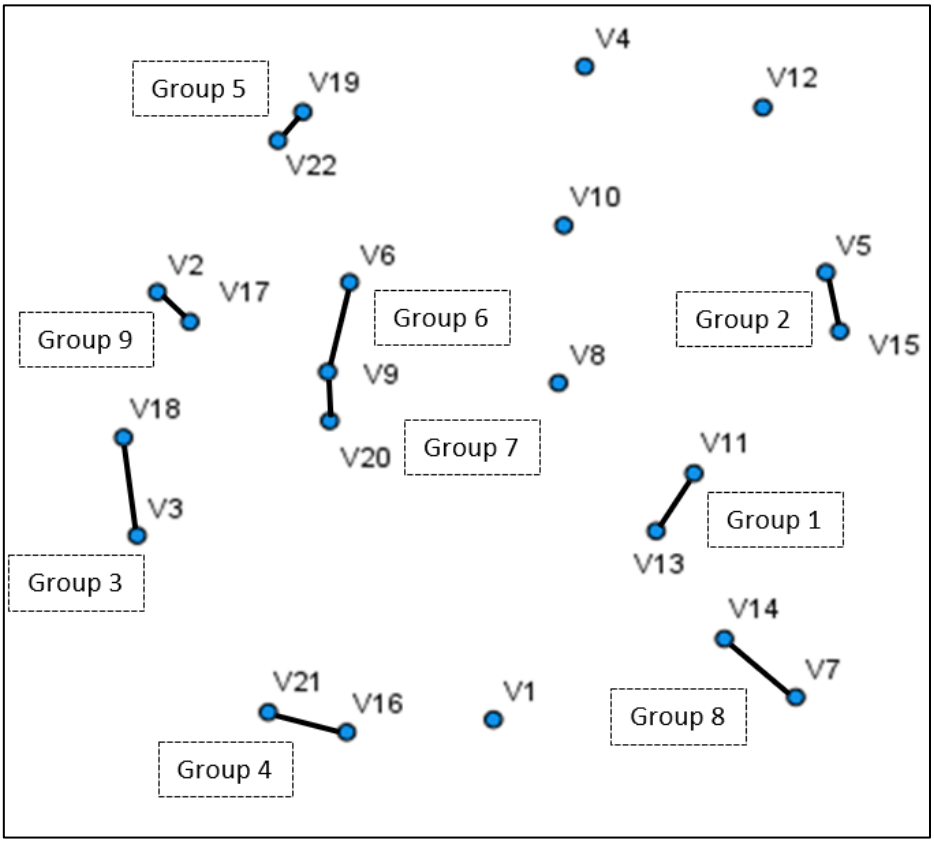

Figure 11. Visualization of research groups by MDS analysis

Furthermore, using a standardized distance of 0.25, nine research groups have been identified. Group 1: Fundamental principles of RL; Group 2: Q-Learning RL algorithms with discrete action spaces; Group 3: Advanced RL algorithms to support continuous action spaces; Group 4: RL algorithms in BEM focusing on HVAC control; Group 5: RL algorithms in HEM focusing on HVAC control; Group 6: Combination of RL and DNN with-in household level; Group 7: Deterministic policy gradient (DPG) control for smart home/building environments; Group 8: Advanced RL techniques for BEM optimization; Group 9: Demand response optimization using RL for HEM.

Table 2. Summary of research groups' key findings and research theme

Research Group	Code	References	Control System/Topic	RL Control Algorithm	RL Agent Type	Control Objectives	Research Theme
Group 1	V11	(Sutton and Barto, 2018)	Physical Tasks (Cart-Pole, Mountain Car), Atari Games.	DQN, DDPG, TRPO	Single Agent	-	Fundamental principles of RL include key components such as states, actions, rewards, and policies. Multiple RL from discrete to continuous algorithms were discussed.
	V13	(Vázquez-Canteli and Nagy, 2019)	Controllable Loads (HVAC), Non-shiftable Loads (Security System), Shiftable Loads (Washing Machine)	Q-Learning, DQN, DDPG, PPO, SAC, MARL	Single/Multi-Agent	Energy Cost, Load Balancing, Peak Shaving, User Comfort	
Group 2	V15	(Watkins and Dayan, 1992)	Gridworld, Maze Navigation, Robotics, Game Playing	Q-Learning	Single Agent	-	The applications of Q-Learning with discrete action spaces. Lu and Hong concentrated on the demand response strategies in smart grids.
	V5	(Lu and Hong, 2019)	Demand Response in Smart Grids	Q-Learning	Single Agent	Profits of Service Providers, Customers	
Group 3	V3	(Lillicrap et al., 2019)	Physic Tasks (Cartpole Swing-up, Dexterous Manipulation, Legged Locomotion, Car Driving)	DPG	Single Agent	-	The applications of advanced RL algorithms to support continuous action spaces.
	V18	(Ye et al., 2020)	Residential multi-energy system (MES) with ESS, TES, EHP, GB, PV	MDP, PDDPG	Single Agent	Energy Cost	
Group 4	V16	(Wei et al., 2017)	Building Energy Management (BEM) focusing on HVAC Control	DQN	Single Agent	Energy Cost, Thermal Comfort	RL algorithms in BEM focusing on HVAC control.
	V21	(Yu et al., 2021b)	BEM focusing on HVAC Control	MAAC	Multi-Agent	Energy Cost, Thermal Comfort	
Group 5	V22	(Zhang et al., 2016)	Home Energy Management (HEM) focusing on HVAC Control	MLP, Regression based Learning	Single Agent	Energy Cost, Thermal Comfort	RL-based HEM mainly focuses on HVAC control to minimize the total electricity cost and thermal discomfort.
	V19	(Yu et al., 2019)	HEM focusing on HVAC Control with PV Integration	LOT	Single Agent	Energy Cost, Thermal Comfort	
Group 6	V6	(Lu et al., 2019)	HEM with Controllable Loads (HVAC), Non-shiftable Loads (Security System), Shiftable Loads	Q_Learning	Multi-Agent	Energy Cost, Thermal Comfort	Combination of deep neural networks (DNNs) for uncertainty prediction and RL for energy consumption schedules in household energy management.
	V9	(Mocanu et al., 2019)	BEM with Controllable Loads (HVAC), Non-shiftable Loads (Security System), Shiftable Loads	DQN, DPG	Single Agent	Energy Cost, Peak Reduction, Thermal Comfort	
Group 7	V9	(Mocanu et al., 2019)	BEM with Controllable Loads (HVAC), Non-shiftable Loads (Security System), Shiftable Loads	DQN, DPG	Single Agent	Energy Cost, Peak Reduction, Thermal Comfort	Deterministic policy gradient (DPG) based single-agent RL control in optimizing energy management within the context of smart environments.
	V20	(Yu et al., 2020)	HEM focusing on HVAC Control with PV Integration	DDPG	Single Agent	Energy Cost, Thermal Comfort	
Group 8	V7	(Mason and Grijalva, 2019)	BEM with the integration of HVAC, EV, BESS	Q-Learning, SARSA, AC	Single Agent	Energy Cost, Energy Flexibility, User Comfort	Advanced control techniques, specifically RL and DRL, to optimize energy management in building environments.
	V14	(Wang and Hong, 2020a)	BEM with different control subjects (HVAC, BESS, TES, Appliances)	PG, AC, Value-Based	Single/Multi-Agent	Energy Cost, Energy Flexibility, User Comfort	
Group 9	V2	(Li et al., 2020)	HEM with Controllable Loads, Non-shiftable Loads, Shiftable Loads, EV	TRPO	Single Agent	Energy Cost, Thermal Comfort	Optimization of HEM through RL techniques in the context of demand response (DR) strategies.
	V17	(Xu et al., 2020)	HEM with Controllable Loads, Non-shiftable Loads, Shiftable Loads, EV, PV	Q-Learning	Multi-Agent	Energy Cost, Thermal Comfort	

Research Group 1: Fundamental principles of RL

First, research group 1 includes two highly related publications (Sutton and Barto, 2018; Vázquez-Canteli and Nagy, 2019) that discuss the applications of RL algorithms in various domains. While Sutton and Barto's book provides a comprehensive overview of RL techniques, Vázquez-Canteli's paper focuses specifically on the role of RL in demand response applications within the smart grid context.

Multiple studies have further developed this research group, especially in the context of multi-energy systems including electricity, heating, and cooling for grid-interactive buildings. Deltetto et al., 2021 have proved that a combination of deep RL and RBC algorithms can reduce energy consumption and energy costs of a cluster of small buildings by 7% and 4% respectively. Another study has examined the effectiveness of an advanced RL algorithm (SAC) in optimizing the operational costs of an office building equipped with integrated energy systems including thermal energy storage, battery storage, and solar PV generation (Brandi et al., 2022). The results indicate that the proposed control strategy significantly reduces operational energy costs compared to the fully rule-based control, achieving savings ranging from 39.5% to 84.3% across various configurations. Different from the above work that uses a single-agent RL-based strategy, Xie et al., (2023) have presented an approach to demand response in grid-interactive buildings by combining a shared attention mechanism and an actor-critic algorithm with multi-agent deep RL (MADRL). By assigning an agent to every building, MADRL facilitates the implementation of decentralized cooperative policies aimed at reducing electricity expenses and optimizing load shaping. The approach's efficacy is demonstrated by computational tests, which yield a reduction in net load demand of more than 6% when compared to conventional RL techniques.

Research Group 2: Q-Learning RL algorithms with discrete action spaces

Research Group 2 focuses on Q-Learning based applications with discrete action spaces on control domains, especially in the home/building energy management context. Watkins and

Dayan, in 1992 laid the groundwork for Q-Learning, presenting the theoretical underpinnings including the proof of convergence and an explanation of the iterative learning process involving states, actions, and rewards. Lu and Hong, in 2019, introduced an innovative demand response algorithm for smart grids, which integrates Q-Learning with deep neural network (DNN) techniques.

Research Group 2's literature branches into two main research directions. One avenue extends the original theme by incorporating prediction algorithms like feedforward neural networks (FFNN) or long-short-term memory (LSTM) neural networks with Q-Learning. This extension aims to address uncertainties such as fluctuating electricity prices and indoor temperatures, ultimately striving to generate optimal electrical demand profiles and offer grid services while ensuring user comfort in smart environments like home and residential energy management (Lu et al., 2019; Ojand and Dagdougui, 2022). The other avenue emphasizes the superiority of advanced algorithms such as soft actor-critic (SAC) and deep reinforcement learning (DRL) over traditional Q-Learning, particularly in handling continuous state and action spaces (Pinto et al., 2021b, 2021a).

Research Group 3: Advanced RL algorithms to support continuous action spaces

Despite Q-Learning's effectiveness in diverse domains like physical control tasks and energy system optimization, it encounters difficulties when confronted with high-dimensional state spaces and continuous action spaces (Lillicrap et al., 2019; Ye et al., 2020). Research Group 3 investigates the use of model-free RL algorithms such as DPG and DDPG, which expand upon Deep Q-Learning to support continuous action spaces. Ye et al., 2020 present a real-time autonomous energy management strategy for residential Multi-Energy Systems (MES) using a model-free deep reinforcement learning (DRL) approach. The study combines the deep deterministic policy gradient (DDPG) algorithm with prioritized experience replay to optimize energy usage while minimizing costs and handling uncertainties such as electricity price and indoor temperature. This approach operates in multi-dimensional continuous state and

action spaces, achieving lower energy costs compared to other DRL methods and traditional optimization techniques.

With a particular focus on improving energy management in home and residential buildings, several academics have evolved Deterministic Policy Gradient (DPG) algorithms into state-of-the-art techniques such as Soft Actor-Critic (SAC), twin delayed DDPG (TD3) and Trust Region Policy Optimization (TRPO). Zenginis et al. developed an RL framework using the TD3 algorithm for real-time energy management in smart homes, integrating PV, ESS, and HVAC systems. Lu et al. 2023 introduced a Reward Shaping (RS)-based Actor-Critic Deep Reinforcement Learning (ACDRL) algorithm designed for managing residential energy consumption profiles amidst uncertain factors. Real-world case studies demonstrate that the proposed algorithm surpasses existing RL methods in terms of learning speed, solution optimality, and cost reduction, achieving approximately 38.57% lower electricity costs compared to cases without scheduling.

Research Group 4: Applications of RL algorithms in HVAC control for BEMs

The two core publications of Research Group 4 focus on the applications of DRL techniques to optimize the operation of building HVAC systems. They acknowledge the effectiveness of DRL-based algorithms in reducing energy costs while maintaining optimal room temperatures. Besides, both studies emphasize the importance of data-driven approaches in controlling building HVAC systems. They leverage real-world data, including weather and pricing data, to develop and validate their algorithms, highlighting the potential of data-driven methods in enhancing building energy efficiency (Wei et al., 2017; Yu et al., 2021b). The successors of the core publications in Research Group 4 have expanded the primary research direction concerning RL algorithms in HVAC control for BEMs into various avenues, considering multiple factors affecting energy consumption of buildings or offices, such as the integration of renewable energy, user occupancy, or multizone control (Deng et al., 2022; Macieira et al., 2021; Shen et al., 2022). For instance, Shen et al., proposed a multi-agent deep RL framework utilizing a dueling double deep Q-network for single-agent optimization and a value-

decomposition network for cooperation optimization among multiple agents in 2022. Key achievements and conclusions of the study include the establishment of mathematical models for BES equipment and load calculation, deploying an intelligent D3QN-PER algorithm for complex control tasks, and applying the VDN algorithm for multi-agent deep RL. Evaluation using established methods reveals substantial improvements in reducing uncomfortable duration, unconsumed renewable energy, and energy costs compared to benchmark control models.

Research Group 5: Applications of RL algorithms in HVAC control for HEMs

Different from Research Group 4 which concentrates on optimization control of HVAC systems for buildings using RL and machine learning methods, Research Group 5 puts effort into improving energy scheduling or management within household environments, addressing the challenges posed by decentralized energy systems and seeking to minimize energy costs while maintaining user satisfaction or comfort. The two core studies in this group account for uncertainties in energy usage behavior, real-time electricity prices, outdoor temperature, renewable generation output, and other relevant factors. They propose algorithms or frameworks capable of handling these uncertainties to achieve optimal energy management outcomes (Yu et al., 2019; Zhang et al., 2016).

It is worth mentioning that while studies in the context of home energy management typically aim to address the complexities of decentralized energy systems, publications focusing on buildings' energy management tend to concentrate on minimizing energy costs associated with HVAC systems in multi-zone structures.

Research Group 6: Combination of DNN-based prediction and RL-based control within household level

Publications forming Research Group 6 highlight the significance of Demand Response (DR) strategies in enhancing the efficiency and stability of power systems, particularly in the realm

of home energy management. RL such as Q-Learning and Deep Q-Network and deep learning techniques are combined for uncertainty prediction and energy consumption schedules with the primary goal of developing robust energy management schemes capable of handling multi-appliance energy management and optimizing user utility while mitigating costs (Lu et al., 2019; Mocanu et al., 2019).

Research Group 7: Deterministic policy gradient (DPG) based single-agent RL control for smart home/building environments

The articles from Research Group 7 highlight the benefits of using DRL algorithms for scheduling energy resources and controlling HVAC and energy storage systems (ESS) in smart homes and buildings (Mocanu et al., 2019; Yu et al., 2020). DPG techniques are used to improve energy management systems. The articles also acknowledge the complexity and uncertainty of energy management systems. These point out problems such as model uncertainty, parameter uncertainty, and temporally coupled operational limitations, highlighting the need for creative algorithms to address these issues successfully.

Research Group 8: Reviews regarding advanced RL techniques for BEM optimization

Mason and Grijalva, 2019; Wang and Hong, 2020 explored the utilization of RL in BEM or building controls, noting RL's potential to enhance building performance and energy efficiency. Both investigations thoroughly examined existing studies or literature concerning RL's implementation in their specific fields. Consequently, they extensively deliberated on the obstacles and impediments hindering the widespread integration of RL in actual building settings. These challenges include time-consuming training processes, concerns about control security and robustness, and limitations in generalization capabilities.

Research Group 9: Demand response optimization using RL for HEM

The research group centers around the optimization of home appliance scheduling and energy consumption to facilitate efficient demand response at the household level. Both studies that form the research group, propose novel frameworks that employ RL algorithms to address the complexities of demand response scheduling. They utilize RL approaches such as DRL and Q-learning to optimize appliance schedules and minimize electricity costs (Li et al., 2020; Xu et al., 2020). Furthermore, the proposed frameworks incorporate neural network-based policies to make decisions based on high-dimensional sensory data and predictions of electricity prices and solar photovoltaic (PV) generation. They leverage NNs for processing real-time data and optimizing energy consumption schedules.

4.5 Conclusion and Future Research Directions

The literature review conducted for the thesis thoroughly analyzed the key studies on the applications of RL algorithms in optimizing building energy management. A comprehensive database of literature was extracted from the Web of Science, comprising 143 highly relevant publications on the subject during the past ten years. Utilizing bibliometric techniques, particularly co-citation analysis, the extracted database was examined, resulting in the identification of nine distinct research groups. Furthermore, the review identified and analyzed the expansion of research streams and directions within each research group:

Group 1: Fundamental principles of RL.

Group 2: Q-Learning RL algorithms with discrete action spaces.

Group 3: Advanced RL algorithms supporting continuous action spaces.

Group 4: RL algorithms in building energy management (BEM) focusing on HVAC control.

Group 5: RL algorithms in home energy management (HEM) focusing on HVAC control.

Group 6: Combination of RL and DNN in EMS at the household level.

Group 7: Deterministic policy gradient (DPG) control for smart home/building environments.

Group 8: Advanced RL techniques for BEM optimization.

Group 9: Demand response optimization using RL for HEM.

In conclusion, RL plays a crucial role in transforming Energy Management Systems within buildings by enabling intelligent, adaptive, and autonomous control of energy resources by:

- **Handling Complex, Dynamic Environments**: RL enables EMS to adapt to the ever-changing environment by learning optimal control policies through continuous interaction with the system. RL algorithms explore different actions (e.g., adjusting HVAC settings, managing battery storage, charging/discharging EV batteries, or controlling renewable energy systems) and learn from the feedback (rewards or penalties) based on energy costs, grid signals, and occupant comfort.
- **Optimizing Energy Efficiency and Cost Savings**: Unlike traditional control approaches that focus on short-term goals (e.g., reducing energy use in the current time step), RL optimizes for long-term rewards. This is particularly important in grid-interactive buildings where actions taken today (such as charging or discharging batteries or pre-cooling a building) have future implications in terms of energy consumption, grid interaction, and occupant comfort.
- **Demand Response and Grid Flexibility**: RL can be used to automate demand response strategies by learning to shift or curtail energy loads during peak grid demand periods, thereby reducing the building's reliance on the grid. By learning from the outcomes of past DR events, RL can improve its ability to participate in grid services while minimizing occupant discomfort.
- **Managing Distributed Energy Resources (DERs)**: RL can learn how to optimally manage on-site renewable energy systems, ensuring that energy generated from solar panels or wind turbines is used effectively. For instance, it can decide when to store excess renewable energy in batteries or when to export it to the grid, based on current and predicted energy needs and prices.
- **Overcoming the Challenges of conventional controls**: RL does not require an explicit model of the building or energy system, making it more adaptable and easier to deploy in complex environments. Instead, RL learns from its interactions with the environment, making decisions based on direct feedback rather than relying on predefined models. This is particularly useful when building systems evolve or when dealing

with highly variable conditions such as renewable energy fluctuations, occupant behaviour changes, or unpredictable EV arrival and departure schedules.

One noteworthy avenue for future research is the evaluation of proposed algorithms in terms of their impact on carbon footprint or greenhouse gas emissions. While many studies have applied RL algorithms with the aim of reducing greenhouse gas emissions from buildings and households, few have included carbon emissions as a key performance indicator in their evaluations.

Additionally, despite the development of numerous simple and complex RL algorithms for various energy systems in buildings and households, there is a lack of comprehensive studies that consider a basic energy system incorporating renewable energy generation, EV chargers, electrical loads (controllable, non-controllable, shiftable, non-shiftable), batteries, and thermal systems. Such a study could compare state-of-the-art RL algorithms (such as RBC, Q-Learning, DQN, and SAC) against multiple key performance indicators for such energy systems.

5 Simulation Problem Formulation

The simulation part of the research will focus solely on the design and implementation of an energy management system utilizing the Soft Actor-Critic RL algorithm as the method supports continuous actions and state spaces that are required to handle the complexity of GEBs including residential loads, renewable energy generation (solar PV), active energy storage system (batteries), utility grids, and EVs with two-way energy transmission (Vehicle-2-Grid, Grid-2-Vehicle).

Figure 12 illustrates the building model that integrates electricity sources to power controllable distributed energy resources (DERs), including electric devices and energy storage systems (ESSs), which are used to meet non-shiftable electrical loads, provide the required energy for EV charging, and offer energy flexibility to the grid. In this model, thermal loads, thermal energy storage systems, and occupant models are excluded.

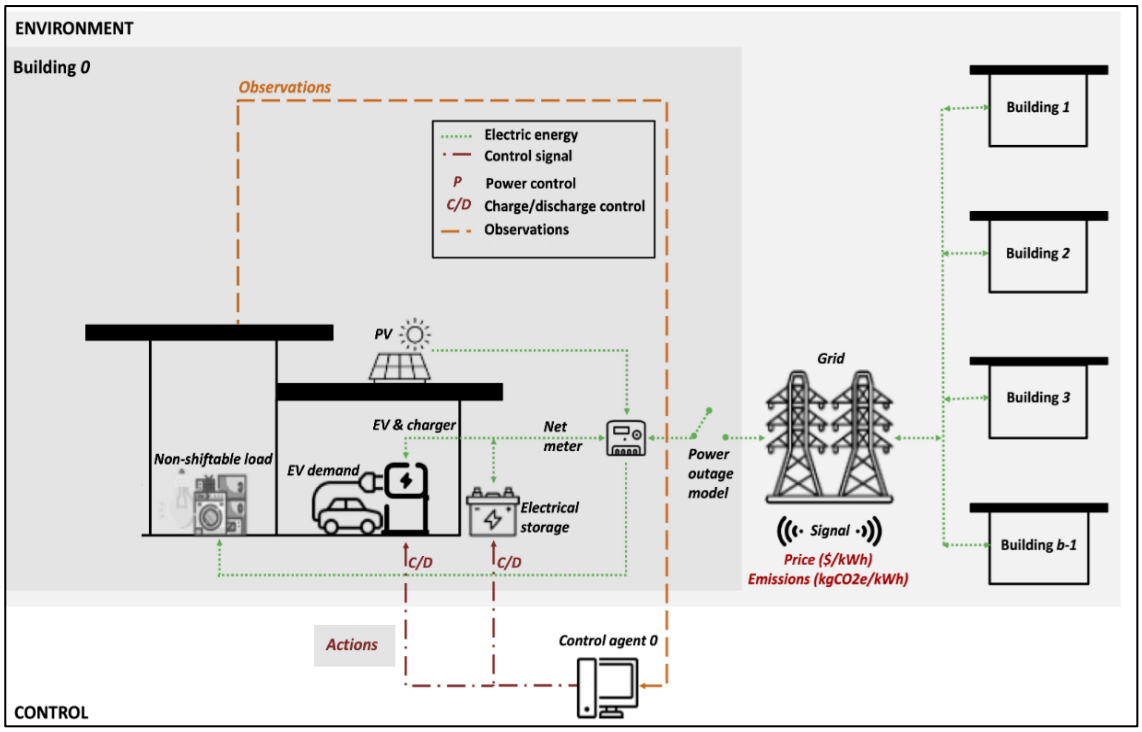

Figure 12. System model including electricity sources, loads, ESSs, and EVs (Nweye et al., 2024).

The main objective of the simulation in this study is to optimize the energy management of a group of buildings by leveraging solar PV generation, Battery Energy Storage Systems (BESS), and EV batteries to maximize energy flexibility while satisfying non-shiftable loads, and EV charging demand. This is achieved through a central control agent using Soft Actor-Critic (SAC) deep RL to coordinate and manage the resources efficiently.

5.1 System Model

The entire system model represents a group of two buildings, where each building consists of a combination of electricity sources powering controllable distributed energy resources (DERs). These include electric devices and energy storage systems (ESSs) that meet electrical demands and provide the grid with energy flexibility.
The system model will not take into account the thermal dynamics model of each building, as well as the possible power outage model of the utility grid.

5.1.1 Loads

The simulation considers two distinct types of loads in a building: non-shiftable loads (electric appliances), and EV loads. Electric equipment refers to non-flexible plug loads like lighting, entertainment systems, security, and monitoring systems appliances. The EV consumption corresponds to the energy needed to charge an EV to its target SoC before a scheduled departure.

5.1.2 Energy Storage Systems

There are two main BESSs in a building including the stationary battery energy storage system (BESS) and the EV BESS. The stationary BESS is a type of DER that powers any electric device in a building when in discharge mode. It can also be charged by one or more of the electricity sources such as the utility grid, PV, or the EV BESS. The surplus energy is exported to the

utility grid as part of the building's net energy if the electrical storage delivers more energy than is required to fulfill the building's demand. The EV BESS is only available on a schedule that is determined by its arrival and departure times, however, the EV is a DER type and functions similarly to the stationary BESS.

5.1.3 Electricity Sources

The utility grid provides the majority of the electricity for the buildings' electric devices, while a separate photovoltaic system supplies self-generated power. The grid or the PV system can be used to charge the stationary BESS and EV BESS. When discharging, they cooperate with the grid and PV system to provide the buildings with electricity. As part of the buildings' net energy export, any excess energy from EV discharge, BESS, or self-generation flows back to the grid.

$$E_t^{building} = E_t^{nonshiftable\_loads} + E_t^{BESS} + E_t^{EV_{Charger}} + E_t^{PV} \qquad (5.1)$$

Equation (5.1) illustrates that the building's net electricity consumption is calculated by adding the positive electricity usage of all electric devices as non-flexible loads, along with the bipolar electricity consumption of the BESS and EV charger, and the negative electricity generation from the PV system.

The district-level net electricity consumption will be measured by the sum of all $E_t^{building}$ from all buildings.

$$E_t^{building} = \sum_{k=0}^{N-1} E_t^{building\ k} \qquad (5.2)$$

5.1.4 Electric Vehicle Charger (EVC) Energy Model

In the actual world, a building, such as a home, workplace, or public area, may have several EV chargers installed. Similarly, in this simulation, a single building can be equipped with more than one simulated charger. Each charger is modeled to replicate an actual charging plug. This setup allows for the simulation of scenarios where multiple EVs need to be managed simultaneously, whether in a home with several EVs, an office building providing chargers for employees, or a public charging hub with multiple charging stations.

Each charger has its own identity with the format as $charger_{b\_i}$, where *b* stands for the building number, and *i* is the order number of the charger installed in the building *b.* When there is an EV connected to the charger, the electricity consumption at a time step *t*, denoted as $E_t^{charger_{b\_i}}$ is a function derived from the control action at the previous time step $a_{t-1}^{charger_{b\_i}}$.

$$E_t^{charger_{b\_i}} = \begin{cases} a_{t-1}^{charger_{b\_i}} \times p^{charger_{b\_i},\ \text{nominal\_charging}}, & a_{t-1}^{charger_{b\_i}} \geq 0 \\ a_{t-1}^{charger_{b\_i}} \times p^{charger_{b\_i},\text{nominal\_discharging}}, & a_{t-1}^{charger_{b\_i}} < 0 \end{cases} \quad (5.3)$$

Where

$$a_{t-1}^{charger_{b\_i}} \in \begin{cases} [-1, 0], & for\ V2G\ mode \\ [\ 0, 1], & for\ G2V\ mode \end{cases} \quad (5.4)$$

The supplied energy can be calculated as a product of the electricity consumption $E_t^{charger_{b\_i}}$ and the charging/discharging efficiency $\eta^{charger_{b\_i}}$.

$$Q_t^{charger_{b\_i}} = \eta^{charger_{b\_i}} \times E_t^{charger_{b\_i}} \quad (5.5)$$

5.1.5 EV Energy Model

The EV model simulates EVs' real-world operation and practical limitations, focusing on their role as a key factor for energy flexibility within the system. In the simulation, EVs can connect to an EV Charger to consume energy in G2V mode, whereas in V2G mode, they discharge energy back to the grid. The connection and disconnection of EVs to chargers follow a

predefined schedule based on a pre-simulated file, which will be explained more in the Simulation section.

The EV's battery system is the core of the EV model which inherits directly all attributions of Stationary Battery from the CityLearn framework. The EV Battery Energy Storage System (EV BESS) takes into account the battery degradation over time $C_t^{\text{EV\_BESS}}$ as the maximum capacity gradually decreases, a roundtrip efficiency η^{EV_BESS} which presents energy losses during charging and discharging circles. Additionally, the EV BESS can generate a maximum input/output power, which is a product of the nominal power $p^{EV\_BESS,nominal}$ and the SoC-power-dependent function $f(SoC_t^{\text{EV\_BESS}})$.

$$P_t^{\text{EV\_BESS,max}} = p^{EV\_BESS,nominal} \times f(SoC_t^{\text{EV\_BESS}}) \tag{5.6}$$

The stored energy within the EV BESS at a time step *t* is calculated by a piecewise function depending on whether the EV is being charged (G2V) or discharged (V2G).
In G2V mode, the EV BESS stored energy can be formalized as:

$$Q_t^{\text{EV\_BESS,G2V}} = min\left(C_t^{\text{EV\_BESS}}, Q_{t-1}^{\text{EV\_BESS}} + min\left(Q_t^{charger_{b\_i}}, P_t^{\text{EV\_BESS,max}}\right) \times \eta^{EV\_BESS}\right) \tag{5.7}$$

The EV BESS stored energy at a time step *t* during charging (G2V) $Q_t^{\text{EV\_BESS,G2V}}$ is the minimum of either the remained energy after degradation $C_t^{\text{EV\_BESS}}$ and the maximum energy coming from the control action applied via the EV charger $min\left(Q_t^{charger_{b\_i}}, P_t^{\text{EV\_BESS,max}}\right) \times \eta^{EV\_BESS}$ plus the initial energy remaining in the BESS at the previous time step *t-1*. Notice that we are not considering the thermal losses due to heat.
In V2G mode, the EV BESS stored energy can be estimated as:

$$\begin{aligned} Q_t^{\text{EV\_BESS,V2G}} = min\Big(&C_0^{\text{EV\_BESS}} \times DoD^{EV\_BESS}, Q_{t-1}^{\text{EV\_BESS}} \\ &+ min\left(Q_t^{charger_{b\_i}}, -P_t^{\text{EV\_BESS,max}}\right) \div \eta^{EV\_BESS}\Big) \end{aligned} \tag{5.8}$$

The equation represents the total energy remaining after accounting for discharging activities, involving two main calculations. First, a depth-of-discharge (DoD) constraint limits the stored energy, making sure the battery is not completely discharged (DoD > 0). Second, it incorporates the energy drawn out for usage, $Q_t^{charger_{b\_i}}$, by the connected EV charger. This discharge is constrained by the system's power output capacity (denoted as $-P_t^{\mathrm{EV\_BESS,max}}$) taking into account the round-trip efficiency, η^{EV_BESS}.

The EV BESS SoC at any time step *t* is calculated by dividing the stored energy by the nominal capacity before any degradation.

$$SoC_t^{EV\_BESS} = \frac{Q_t^{EV\_BESS}}{C_0^{EV\_BESS}} \tag{5.9}$$

5.2 Intelligent Energy Management System (iEMS) Design

The simulation section is centered on designing and assessing the performance of an iEMS, as illustrated in Figure 13. The system consists of four primary layers: the Data Processing Layer, the Forecasting Layer, the Reinforcement Learning Layer, and the Controllable Devices Layer.

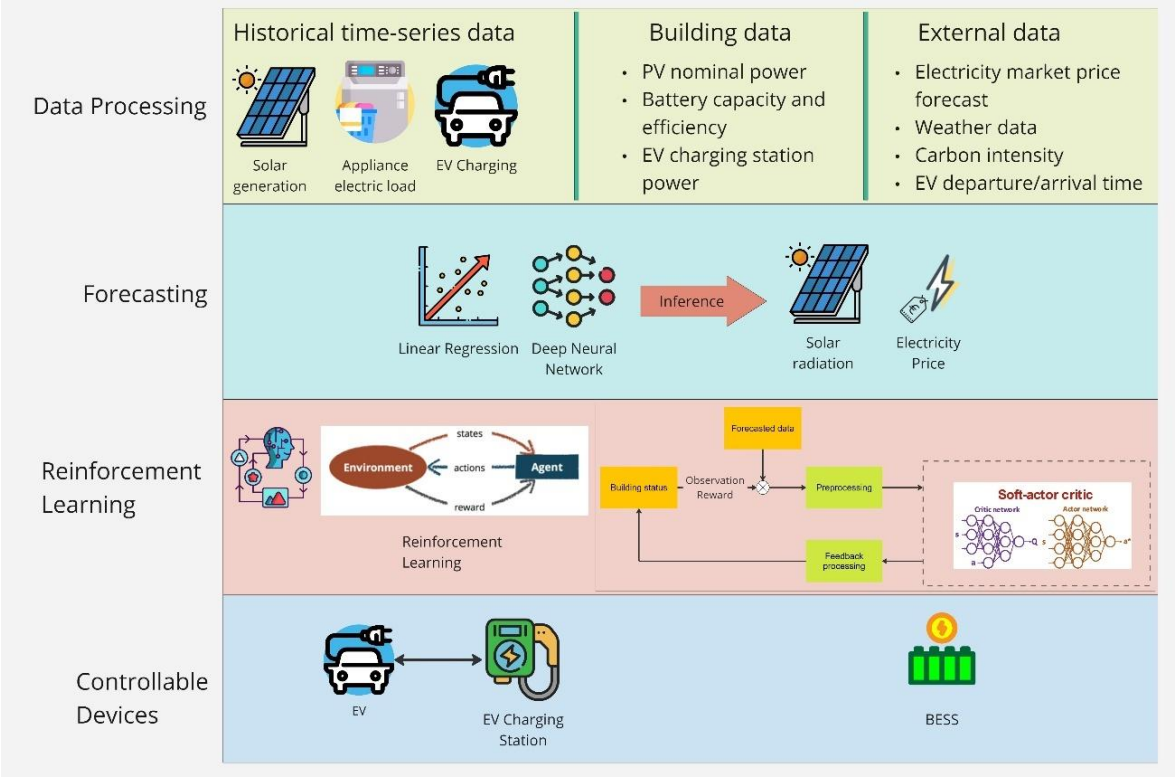

Figure 13. iEMS Design.

5.2.1 Data Processing Layer

The required data for the system is categorized into three main types historical time-series data, building configuration data, and external data. The system relies on historical time-series data, including:

- Solar generation: The amount of energy generated by solar PV over time.
- Appliance electric load: The power consumption patterns of appliances in the building.
- EV arrival and departure schedule: Information regarding the arrival and departure time of EVs, and the expected SoC level at departure.

Building data includes:

- PV nominal power: The rated output power of the building's solar photovoltaic system.
- Battery capacity and efficiency: The energy storage system's performance and capacity to store energy efficiently.

- EV charging station power: The power consumption or generation related to EV charging stations.

External data comes from outside sources and includes:

- Electricity market price forecast: Predicted fluctuations in electricity prices.
- Carbon intensity: Information on the environmental impact of electricity usage in terms of carbon emissions.

5.2.2 Forecasting Layer

The forecasting section predicts solar radiation in the next 6, 12, and 24 hours and markets electricity prices based on the processed data. In this research the following algorithms are employed:

- Linear Regression: A simple statistical method for forecasting.
- Deep Neural Network (DNN): A more advanced machine learning model used for predicting complex patterns, such as electricity market price, solar generation, and appliance load.

The average values of both inference models then, are calculated to make predictions regarding solar radiation and electricity price.

5.2.3 Reinforcement Layer

The layer contains a deep RL model that takes into use the data from the Forecasting Layer, observation data, and calculated rewards from the building environments to calculate the optimized actions that the controllable layer should act to achieve the KPIs. Details regarding the deep RL will be explained later in this chapter.

5.3 Key Performance Indicators (KPIs)

We assess the optimal control performance using multiple KPIs targeted for minimization: electricity bills, carbon emissions, ramping, average daily peak, and (1 - load factor) (Vazquez-Canteli et al., 2020). Another KPI called EV charging satisfaction rate (CSR), is added to assess how well the iEMS meets the charging needs of EV users based on their expected schedules and desired charge levels. Among these KPIs, average daily peak, ramping, and (1 - load factor) are measured at the district level by aggregating hourly net electricity consumption (kWh), $E_h^{district}$ of all buildings in the district. Electricity cost and carbon emissions are calculated at the building level using buildings' hourly net electricity consumption (kWh), $E_h^{building}$.

5.3.1 Electricity Cost

Electricity cost refers to the total cost associated with the imported electricity from the grid used at the building level. It is important to note that the imported electricity cost $\left(E_h^{building} * T_{hour}\right)$ can be negative, as excess energy generated by EVs and solar PV systems can be sold back to the grid. T_{hour} is the electricity rate at time step *hour.*

$$cost = \sum_{h=0}^{n-1} \max\left(0, E_h^{building} * T_{hour}\right) \quad (5.10)$$

5.3.2 Carbon Emissions

The carbon emissions indicator is the sum of building-level carbon emissions (Kg_{CO2e}), $E_h^{building} * O_{hour}$. O_{hour} is the carbon intensity $\left(\frac{Kg_{CO2e}}{hWh}\right)$. Similarly to the electricity cost, the carbon emissions from the build at the time step *hour* can also be negative depending on the carbon intensity at that time step.

$$carbon\ emissions = \sum_{h=0}^{n-1} \max\left(0, E_h^{building} * O_{hour}\right) \quad (5.11)$$

5.3.3 Average Daily Peak

The indicator is defined as the average value of the daily $E_{hour}^{district}$ peak where *d* represents the day index and *n* is the total number of days.

$$average\ daily\ peak = \frac{1}{n}\sum_{d=0}^{n-1}\sum_{h=0}^{23}\max\left(E_{24d+h}^{district}, \dots, E_{24d+23}^{district}\right) \qquad (5.12)$$

5.3.4 Ramping

Ramping refers to the absolute difference between consecutive district electricity loads, measuring the change in grid demand from one time step to the next. It reflects how smoothly the district's energy consumption profile transitions over time. Low ramping indicates a gradual increase in grid demand, even when self-generation (like solar PV) is no longer available in the evening or early morning, which helps maintain stability in the energy supply. High ramping, on the other hand, represents sudden, sharp changes in demand, which can strain the grid infrastructure and increase the risk of blackouts due to a supply-demand imbalance.

$$ramping = \sum_{h=0}^{n-1}\left|E_{h}^{district} - E_{h+1}^{district}\right| \qquad (5.13)$$

5.3.5 Load Factor

The load factor is the ratio of the district's average monthly electricity consumption to its peak electricity demand. In this context, *m* is the month index, *d* is the number of days in a month, and *n* represents the number of months. This metric reflects the efficiency of

electricity usage, with values ranging from 0 (indicating poor efficiency with high peaks and low average usage) to 1 (indicating high efficiency with steady consumption near the peak). The objective is to maximize the $load\ factor$ (minimize $1 - load\ factor$), promoting more consistent and efficient electricity use.

$$1 - load\ factor = \left(\sum_{m=0}^{n-1} 1 - \frac{\left(\sum_{h=0}^{d-1} E_{d.m+h}^{district} \right) \div d}{\max\left(E_{d.m}^{district}, \dots, E_{d.m++d-1}^{district} \right)} \right) \div n \qquad (5.14)$$

5.3.6 EV Charging Satisfaction Rate

CSR is the percentage of the number of times that EVs achieve their target state of charge (SoC) by their expected departure time measured at the building level. It measures the system's ability to meet user charging preferences without compromising mobility or energy needs.

$$EV\ CSR = \frac{Number\ of\ times\ EVs\ meeting\ target\ SoC}{Total\ number\ of\ departure\ times} \times 100 \qquad (5.15)$$

This KPI directly reflects the comfort and convenience for EV users, ensuring the iEMS balances grid efficiency with individual charging requirements. Maximise the CSR, ensuring that the iEMS efficiently manages energy while prioritizing user satisfaction and vehicle readiness.

5.4 Reinforcement Learning

In this section, we introduce the theoretical framework of the RL algorithm. First, a mathematical theory that helps formalize RL problems known as the Markov Decision Process (MDP) is introduced, and then we explain how to implement the chosen Soft Actor-Critic Deep RL for our control problems in EMS.

5.4.1 Markov Decision Process Formulation

An RL problem can be formalized into a Markov Decision Process (MDP), a discrete stochastic control process (Sutton and Barto, 2018).

The EMS problem is converted to an MDP which is a quintuple (S, A, P, R, γ), where:

- S is a set of all observable states representing the controlled environment, called the state space. The state space can be either discrete or continuous. The Markov Property (Bellman, 2010) requires that all states of the environment must be observable at any time step t.
- A is a set of actions, called action space that is available at the state S. Similarly to the state space, the action space may be discrete or continuous. During the state transition, the agent selects the next action based on a specific probability distribution, known as the policy $\pi: S \rightarrow A$. The sequence of states and actions generated through the interaction is denoted as τ: $(s_0, a_0, s_1, a_1, \dots, a_{t-1}, s_t)$. The action represents the decision made by the agent within the control environment to maximize its objectives, as mathematically defined by the reward function.
- P is a state transition probability matrix.

$$P_{ss'}^{a} = P[S_{t+1} = s' \mid S_t = s,\ A_t = a] = P: S \times A \times S' \tag{5.15}$$

  This defines the probability of each possible next state s', given any state s and taken action a. Based on the Markov Property (Bellman, 2010), these probabilities are determined solely by the value of the current state s and do not rely on any prior states of the environment.
- $P: S \times A \times S' \rightarrow R$ is the intermediate reward, received after the transition from state s to state s' due to action a.

$$R_{ss'}^{a} = E[R_{t+1} \mid S_t = s,\ A_t = a, S_{t+1} = s'] \tag{5.16}$$

- $\gamma \in [0,1]$ is the discount factor whose purpose is to reduce the effect of future rewards on the present reward. When $\gamma = 1$, the agent focuses on future rewards rather than immediate ones. Conversely, when $\gamma = 0$, the agent places greater importance on states that provide high immediate rewards.

The control problem can be described using two interconnected value functions: the state-value function, symbolized as $V^{\pi}(s)$, and the action-value function $Q^{\pi}(s,a)$, as presented below:

$$V^{\pi}(s) = E_{\pi}[\sum_{k=0}^{\infty} \gamma^{k} R_{t+k+1} \mid S_t = s] \tag{5.17}$$

$$Q^{\pi}(s) = E_{\pi}[\sum_{k=0}^{\infty} \gamma^{k} R_{t+k+1} \mid S_t = s, A_t = a] \tag{5.18}$$

These functions indicate, respectively, the desirability of being in a particular state S_t in relation to the control objectives and the benefit of performing a certain action A_t in a given state S_t, while following a particular control policy π (Sutton and Barto, 2018).

5.4.2 Soft Actor-Critic Deep Reinforcement Learning

5.4.2.1 Key Characteristics of Soft Actor-Critic

The Soft Actor-Critic (SAC) algorithm leverages an actor-critic architecture, utilizing two separate deep neural networks to approximate the state-value function and the action-value function (Haarnoja et al., 2018).

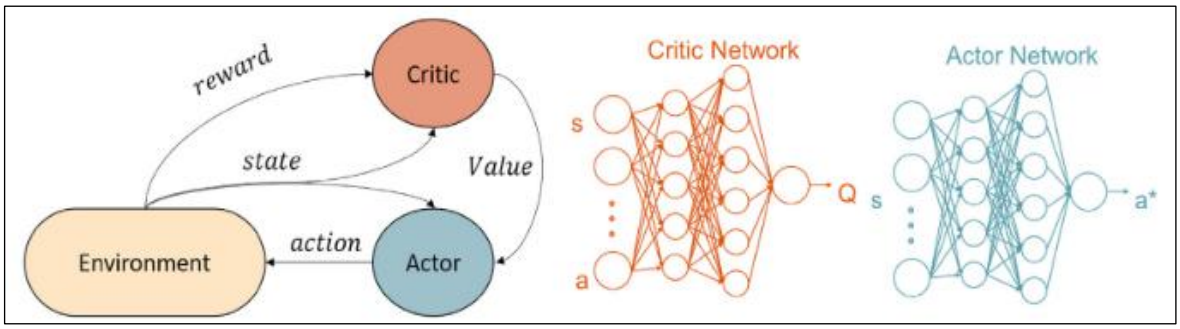

Figure 14. Actor-Critic Environment Interaction and ANNs in SAC (Pinto et al., 2021b).

The key characteristics of a soft actor-critic deep RL algorithm are illustrated in Figure 14 and explained below (Haarnoja et al., 2018):

- **Actor Network (Policy Network)**

  The Actor Network is an ANN which is responsible for choosing the next action a^* based on the current state s of the environment. In SAC, the actor is trained to select actions that maximize rewards and policy entropy, encouraging exploration and avoiding premature convergence to suboptimal solutions. The Actor Network is also known as the Policy Network since it follows the policy gradient method to update the policy parameters in the direction suggested by the critic network.

- **Critic Network (Value Networks)**

  SAC uses two ANNs, called Critic Networks to approximate the action-value function (Q-values). These networks approximate the action-value function (Q-values), which estimates the expected future rewards of taking a particular action in a given state. Having two critics helps reduce overestimation bias and makes the training more stable by using the minimum of the two Q-values during updates.

- **Entropy Regularisation**

  The central feature of SAC is entropy maximization, which encourages the agent to maintain randomness in its actions (exploration) while learning. By optimizing for maximum entropy, the agent is encouraged to explore different actions, rather than always choosing the most certain or best-known action. The balance between reward

maximization and entropy maximization is controlled by an entropy coefficient (α). A higher α means more exploration (higher entropy), while a lower α focuses more on reward maximization (exploitation).

- **Replay Buffer**

  SAC uses a replay buffer to store past experiences (state, action, reward, next state). This allows the algorithm to reuse and learn from previous data, making training more data-efficient. The stored experiences are sampled randomly during the learning process to update the networks.

- **Off-Policy Learning**

  SAC is an off-policy algorithm, meaning it can learn from experiences not necessarily generated by the current policy (actor). This is why the replay buffer is essential, as it allows SAC to update the policy using previously collected experiences.

- **Soft Q-Value Update**

  The Q-value updates in SAC include a soft target, meaning that the update also considers the entropy of the action, not just the expected future reward. This allows the agent to balance between immediate rewards and exploring diverse actions.

5.4.2.2 Entropy Regularised Reinforcement Learning in SAC

Remember the optimal Bellman equation for all standard RL to maximize the expected sum of rewards as:

$$\pi^* = \underset{\pi}{argmax}\ \underset{\tau\sim\pi}{E}\left[\sum_{t=0}^{\infty}\gamma^t\big(R(s_t, a_t, s_{t+1})\big)\right], \qquad (5.19)$$

Where π^* is the optimal policy that maximizes the expected cumulative reward over time and defines the agent's behaviour, specifically how it chooses actions based on the current state; $\underset{\pi}{argmax}$ operation selects the policy that yields the highest cumulative reward; $\underset{\tau\sim\pi}{E}$ represents the expectation (average) over all possible trajectories τ that the agent might take

while following the policy π; $\tau \sim \pi$ means the trajectory τ (a sequence of states and actions $s_0, a_0, s_1, a_1, \ldots$) is generated according to the agent's policy π; $\gamma^t\big(R(s_t, a_t, s_{t+1})\big)$ including the discount factor γ^t, is the cumulative reward the agent expects to receive starting from the time $t = 0$ and continuing indefinitely.

Entropy regularization is the key feature of SAC, where the policy is trained to balance between maximizing the expected return and maintaining entropy, which represents the randomness in action (Haarnoja et al., 2019). This directly relates to the exploration-exploitation trade-off: higher entropy encourages greater exploration, helping the agent learn more effectively in the long run. Additionally, it prevents the policy from settling too quickly into a suboptimal solution.

The entropy $H(P)$ of a probability distribution P (in this case, the policy of the agent) is mathematically defined as:

$$H(P) = \underset{x \sim P}{E}[-logP(x)], \tag{5.20}$$

Where: $P(x)$ represents the probability of taking action x according to the agent's current policy, $\underset{x \sim P}{E}$ is the expected value over the possible actions sampled from the policy, and is the logarithm of the probability of taking action x.

Entropy $H(P)$ measures the uncertainty or randomness of the policy. The higher the entropy, the more uncertain or exploratory the policy is, meaning the agent is more likely to try a variety of actions. The term $-logP(x)$ increases as the probability $P(x)$ decreases, meaning that less likely actions contribute more to the entropy. This encourages exploration by penalizing highly confident or deterministic choices and favouring more diverse, uncertain action choices. This adjustment transforms the RL problem into:

$$\pi^* = \underset{\pi}{argmax}\, \underset{\tau \sim \pi}{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(R(s_t, a_t, s_{t+1}) + \alpha H\big(\pi(\cdot\,|s_t)\big)\Big)\right] \tag{5.21}$$

Where the entropy regularisation $\alpha H\big(\pi(\cdot\,|s_t)\big)$ has been added.

The state-value function V^{π} and action-value function Q^{π} of the MDP can be rewritten with entropy regularisation added (Haarnoja et al., 2018):

$$V^{\pi}(s) = \underset{\tau\sim\pi}{\mathrm{E}}\left[\sum_{t=0}^{\infty}\gamma^{t}\Big(R(s_t,a_t,s_{t+1}) + \alpha H\left(\pi(\cdot|s_t)\right)\Big)\,\middle|\, s_0 = s\right] \qquad (5.22)$$

$$Q^{\pi}(s,a) = \underset{\tau\sim\pi}{\mathrm{E}}\left[\sum_{t=0}^{\infty}\gamma^{t}R(s_t,a_t,s_{t+1}) + \alpha\sum_{t=1}^{\infty}\gamma^{t}H\left(\pi(\cdot|s_t)\right)\,\middle|\, s_0 = s, a_0 = a\right] \qquad (5.23)$$

Where $\underset{\tau\sim\pi}{E}[\cdot]$ is the expectation of the sum of future rewards being averaged over all possible outcomes following the policy π; $R(s_t,a_t,s_{t+1})$ is the immediate reward obtained by taking action a_t in state s_t and transitioning to the next state s_{t+1}; $\alpha H\big(\pi(\cdot\,|s_t)\big)$ is the entropy of the policy π at the state s_t.

SAC sets up the Mean Squared Bellman Error (MSBE) as the loss function used for each critic network. The goal of the MSBE loss is to minimize the difference between the current Q-value estimates and the target Q-values (based on the Bellman equation). This is key to ensuring that the Q-functions accurately estimate the expected return of state-action pairs.

For each Q-function $Q_{\theta_i}(s,a)$, the MSBE loss is calculated as (Haarnoja et al., 2018):

$$L(\emptyset_i, D) = E_{(s,a,r,s',d)\sim D}\left[\Big(Q_{\emptyset_i}(s,a) - y(r,s',d)\Big)^2\right] \qquad (5.24)$$

Where $Q_{\emptyset_i}(s,a)$ is the Q-value estimated by the i-th Q-function (either Q1 or Q2 for the state s and action a; $y(r,s',d)$ is the target Q-value calculated using the Bellman backup:

$$y(r,s',d) = r + \gamma(1-d)\left(\min_{i=1,2} Q_{\emptyset_i}(s',a') - \alpha log\pi(a'|s')\right) \qquad (5.25)$$

This target includes the immediate reward r, the discounted future reward using the minimum of the two Q-functions (to reduce overestimation bias), and the entropy regularisation term $\alpha log\pi(a'|s')$ to encourage exploration. Note that d is just the terminal indicator which receives a value of 0 or 1 to indicate whether the training episode has ended.

Figure 15 below outlines the most important steps to implement the SAC algorithm based on the OpenAI library (OpenAI, 2020). It is noticeable that SAC uses two Q-networks $Q_{\emptyset_i}(s,a), where\ i\ \in [1,2]$ to compute Q-values, and it uses the minimum of the two as the target. This approach, known as clipped double Q-learning, helps mitigate the overestimation bias commonly encountered in RL. Besides, SAC relies on an experience replay buffer *D* to store and randomly sample past experiences (s,a,r,s',d), which helps to stabilize training by reusing past transitions.

Algorithm 1 Soft Actor-Critic

Input: initial policy parameters θ, Q-function parameters ϕ_1, ϕ_2, empty replay buffer $\mathcal{D}$
Set target parameters equal to main parameters $\phi_{\text{targ},1} \leftarrow \phi_1$, $\phi_{\text{targ},2} \leftarrow \phi_2$
repeat
  Observe state s and select action $a \sim \pi_\theta(\cdot|s)$
  Execute a in the environment
  Observe next state s', reward r, and done signal d to indicate whether s' is terminal
  Store (s, a, r, s', d) in replay buffer $\mathcal{D}$
  If s' is terminal, reset environment state.
  if it's time to update **then**
    for j in range(however many updates) **do**
      Randomly sample a batch of transitions, $B = \{(s, a, r, s', d)\}$ from $\mathcal{D}$
      Compute targets for the Q functions:

$$y(r, s', d) = r + \gamma(1 - d)\left(\min_{i=1,2} Q_{\phi_{\text{targ},i}}(s', \tilde{a}') - \alpha \log \pi_\theta(\tilde{a}'|s')\right), \quad \tilde{a}' \sim \pi_\theta(\cdot|s')$$

      Update Q-functions by one step of gradient descent using

$$\nabla_{\phi_i} \frac{1}{|B|} \sum_{(s,a,r,s',d) \in B} \left(Q_{\phi_i}(s, a) - y(r, s', d)\right)^2 \qquad \text{for } i = 1, 2$$

      Update policy by one step of gradient ascent using

$$\nabla_\theta \frac{1}{|B|} \sum_{s \in B} \Big(\min_{i=1,2} Q_{\phi_i}(s, \tilde{a}_\theta(s)) - \alpha \log \pi_\theta \left(\tilde{a}_\theta(s)| \, s\right) \Big),$$

      where $\tilde{a}_\theta(s)$ is a sample from $\pi_\theta(\cdot|s)$ which is differentiable wrt θ via the reparametrization trick.
      Update target networks with

$$\phi_{\text{targ},i} \leftarrow \rho \phi_{\text{targ},i} + (1 - \rho)\phi_i \qquad \text{for } i = 1, 2$$

    end for
  end if
until convergence

Figure 15. Soft Actor-Critic algorithm (OpenAI, 2020).

6 Simulation

In this project, we implement a Soft Actor-Critic (SAC) deep RL algorithm to optimize energy management for a group of buildings using the CityLearn environment, a platform designed for evaluating energy management strategies in urban settings. The simulation is extended to incorporate EVs and EV chargers, adding complexity to the management of energy systems. The integration of EVs introduces additional layers of dynamic energy demand and storage, further challenging the optimization process.

SAC's capacity to balance exploration and exploitation, while effectively handling continuous state and action spaces, makes it ideal for this task. The extended environment simulates realistic interactions between buildings, the grid, renewable energy sources, and EV infrastructure, offering a rich testbed for assessing the algorithm's performance in optimizing energy use, reducing peak demand, and maximizing energy flexibility.

This setup aims to demonstrate SAC's potential in managing complex, dynamic energy systems efficiently, providing insights into future smart city implementations with sustainable and intelligent energy solutions.

6.1 Simulation Environment

CityLearn is a simulation environment designed for the study and development of energy management solutions in urban settings (Nweye et al., 2024). It provides a framework to model and simulate the behaviour of multiple buildings within a city, focusing on the optimization of energy usage, storage, and distribution. By incorporating elements such as building energy consumption, renewable energy generation, and energy storage systems, CityLearn enables researchers and developers to test and evaluate RL algorithms and other control strategies for improving the energy efficiency and sustainability of urban areas. This environment is particularly useful for exploring how decentralized energy systems can contribute to smart grid applications and demand response scenarios.

In this simulation project, we will utilize the following existing implementation of the City-Learn Framework:

- Electricity Source: Grid, Solar PV
- Energy Storage: Battery
- Electric Load: Non-shiftable load

Note that for the focus on EVs and the simplicity of the energy system, we will not take into account the HVAC, heat pump, and DHW energy storage systems.

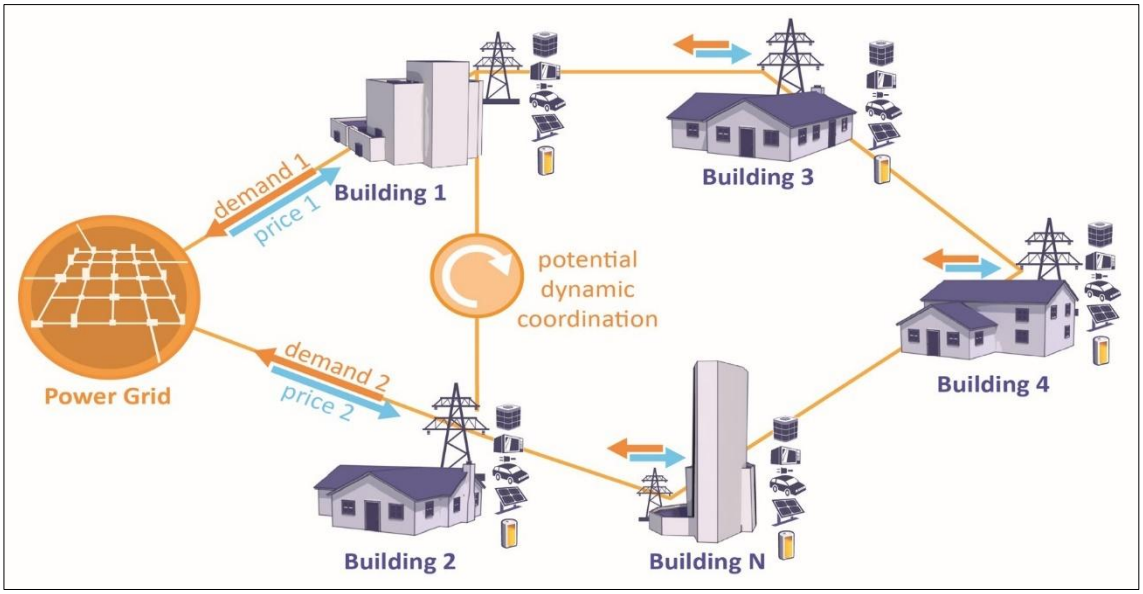

Figure 16. An overview of a district energy system in CityLearn (Vázquez-Canteli et al., 2019).

6.1.1 Dataset Descriptions

The dataset used in the simulation, *citylearn_challenge_2022_phase_all*, is derived from 17 zero net energy (ZNE) single-family homes in the Sierra Crest Zero Net Energy community in Fontana, California. These buildings were analyzed for grid integration within ZNE communities as part of the California Solar Initiative program, focusing on the impact of high PV generation penetration and on-site electricity storage. The dataset spans one year from August 1, 2016, to July 31, 2017. In the constructed community, eight of the 17 homes feature batteries with a 6.4 kWh capacity, a 5 kW power rating, 90% round-trip efficiency, and a 75% depth of discharge. Homes have an installed PV capacity varying from 4 kW to 15 kW (Vázquez-Canteli et al., 2019). In this simulation, we randomly select 2 buildings which are

Building 1 and Building 7, as they both have solar generation installed and have different patterns of electrical loads.

We extend the original dataset, *citylearn_challenge_2022_phase_all*, to include data for a set of EVs that are supposed to connect to EV chargers installed in the buildings. The data for EVs was created using the simulation software SimBev 2.0.0 (Reiner, 2024) that can generate forecasts of charging demand for EVs in a given period. We focus specifically on a subset of the generated data that includes key details such as the Estimated Departure Time, Required SOC at Departure, Estimated Arrival Time, and Estimated SOC at Arrival.

Table 3. Energy system properties of target buildings.

Building No	**PV [KW]**	**Battery**			**EV Chargers**		**EV**	
		Capacity [kWh]	Power [kW]	Efficiency	Power [kW]	Efficiency	Capacity [kWh]	Power [kW]
1	12	6.4	5	0.9	11	0.95	60	50
7	9	6.4	5	0.9	7.4	0.90	50	50

The electricity rate plan employed in the study is derived from CityLearn's dataset city-learn_challenge_2022_phase_all, which was based on the TOU-D-PRIME rate offered by Southern California Edison, the utility provider for the community. This rate plan, outlined in Table 4, is tailored for customers with residential batteries (Southern California Edison, 2022).

Table 4. Electricity rate ($/kWh) in the simulation environment.

Time	**June - September**		**October - May**	
	Weekday	Weekend	Weekday	Weekend
08:00 – 16:00	0.21	0.21	0.20	0.20
16:00 – 21:00	0.54	0.40	0.50	0.50
21:00 – 08:00	0.21	0.21	0.20	0.20

Figure 17 illustrates that under this plan, electricity is most affordable during the early morning and late night hours, with reduced rates during off-peak months (October to May). Additionally, weekend rates offer lower prices during the peak hours of 16:00 to 21:00.

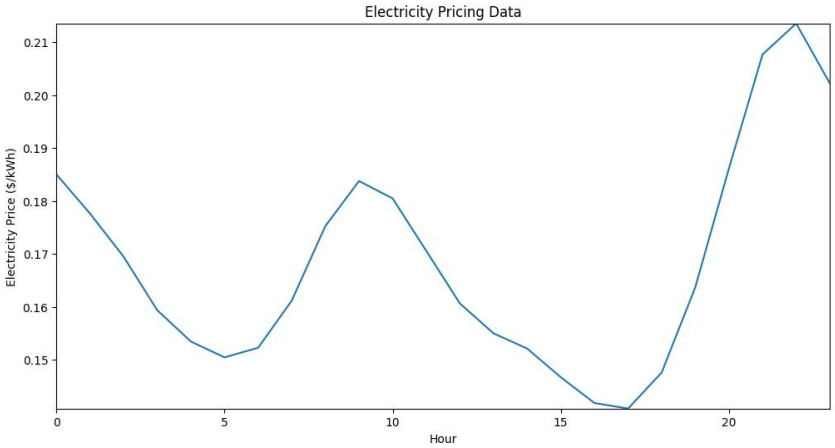

Figure 17 Daily average electricity rate time series data.

Figure 18 highlights the daily average non-shiftable loads and solar generation of the target buildings 1 and 7 used in the simulation. For Building 7, the non-shiftable loads hit a noticeable peak at around 10:00, reaching just over 2.5 kWh. In contrast, Building 1 experiences multiple peak loads at approximately 6:00, 10:00, and 19:00, with values of 1.1 kWh, 1.5 kWh, and a maximum of 2.0 kWh, respectively. On the other hand, solar generation for both buildings looks quite similar, peaking around noon. There is no solar generation before 5:00 and after 20:00, suggesting these time steps correspond to nighttime or non-sunny periods. The solar generation ramps up significantly after 5:00, peaks at noon, and gradually declines toward 20:00.

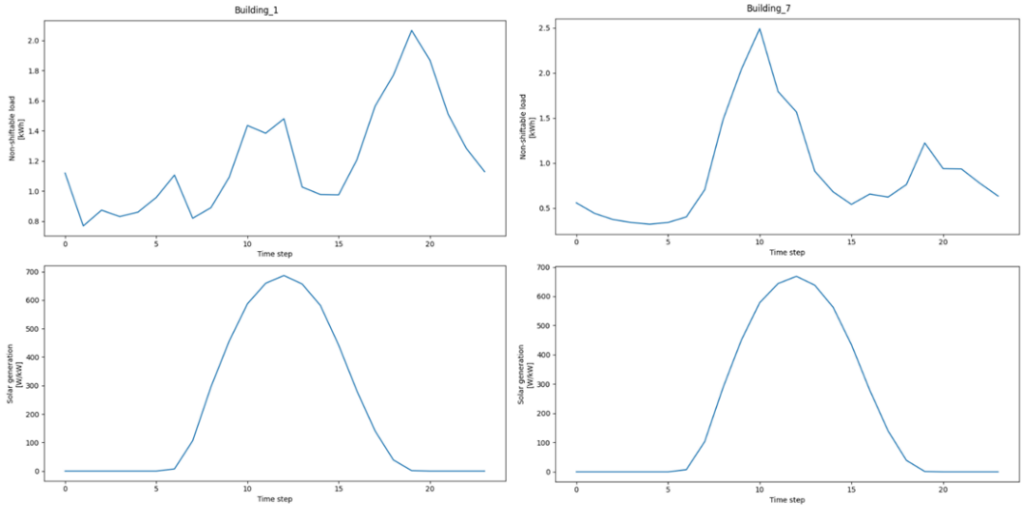

Figure 18. Daily average non-shiftable loads and solar generation of buildings 1 and 7.

6.1.2 Extending CityLearn Environment with EVs Data

As of version 2.2b, CityLearn 2.2b does not provide support for EV-related functionalities, such as simulating commuting patterns or daily routines of EVs. Additionally, it does not model EV-specific behaviours like energy consumption, charging, or disconnection events that occur as vehicles move between locations. To overcome this challenge, we utilize SimBev 2.0.0 (Simulation of Battery Electric Vehicles) (Reiner, 2024) as a data generation tool to provide CityLearn with essential simulated EV data that captures realistic commuting, energy use, and charging behaviours.

SimBev can simulate the daily travel behaviours of EV users, including departure and arrival times, trip lengths, and driving habits. These patterns can be adjusted based on factors such as geographic location, user preferences, and infrastructure (charging station availability).
The output of SimBev consists of a set of files, one for each EV in the scenario. These files are organized as a series of arrays, with each array representing a different aspect of the EVs' operation at a specific day and time, as follows:

TimeStamp:

A time series of months and hours is provided, with values ranging from 1 to 12 for months (January to December) and from 1 to 24 for hours. When integrating with CityLearn, these time references align with the simulation timeline in CityLearn.

EV State:

This indicates the EV's status, with values representing whether it is plugged in and ready to charge (1), approaching a charger but not yet connected (2), or in transit and not connected (3). The EV State helps determine when the EV is available for charging or vehicle-to-grid (V2G) services and when it is consuming battery power during travel. This attribute is useful for optimizing charging schedules based on the EV's operational state.

EV Charger:

This specifies the charging station is connected to or planning to connect to. It holds no specific value ('nan') if no destination charger is specified or a charger ID.

Estimated Departure Time:

This is the anticipated time when the EV is expected to leave its current location (e.g., home, workplace, or charging station). Knowing the estimated departure time allows energy management systems to optimize charging schedules to avoid peak grid times or to charge during times of high renewable energy availability.

Required SOC at Departure:

The "Required SOC at Departure" indicates the minimum battery charge needed to meet the next trip's energy demands. For example, if a long commute is anticipated, a higher SOC may be required. This attribute is essential for ensuring that the EV has adequate charge for the planned journey, without unnecessary overcharging.

Estimated Arrival Time:

This is the predicted time when the EV will arrive back at a location where it can be parked and possibly recharged i.e., home. Knowing the arrival time helps in planning the EV's availability for charging or vehicle-to-grid (V2G) services, as well as understanding when the EV will be idle and able to contribute to energy flexibility.

Estimated SOC at Arrival:

This represents the expected battery level (SOC) when the EV arrives. It accounts for the energy consumed during the journey and helps determine how much additional charging will be needed before the next trip. By knowing the SOC at arrival, energy management systems can decide if charging is immediately necessary or if it can be delayed to align with lower-cost or lower-emission electricity availability.

6.2 Simulation Implementation

6.2.1 Extending CityLearn with EVs Implementation

The simulation environment is built on the foundational structure of CityLearn, which is based on the OpenAI Gym standard (Brockman et al., 2016). Consequently, the core implementation revolves around the *Environment* class that inherits directly from the Gym. The domain model of the simulation platform comprises various components, including buildings, heat pumps, electric heaters, thermal storage tanks, stationary batteries, and PV systems.

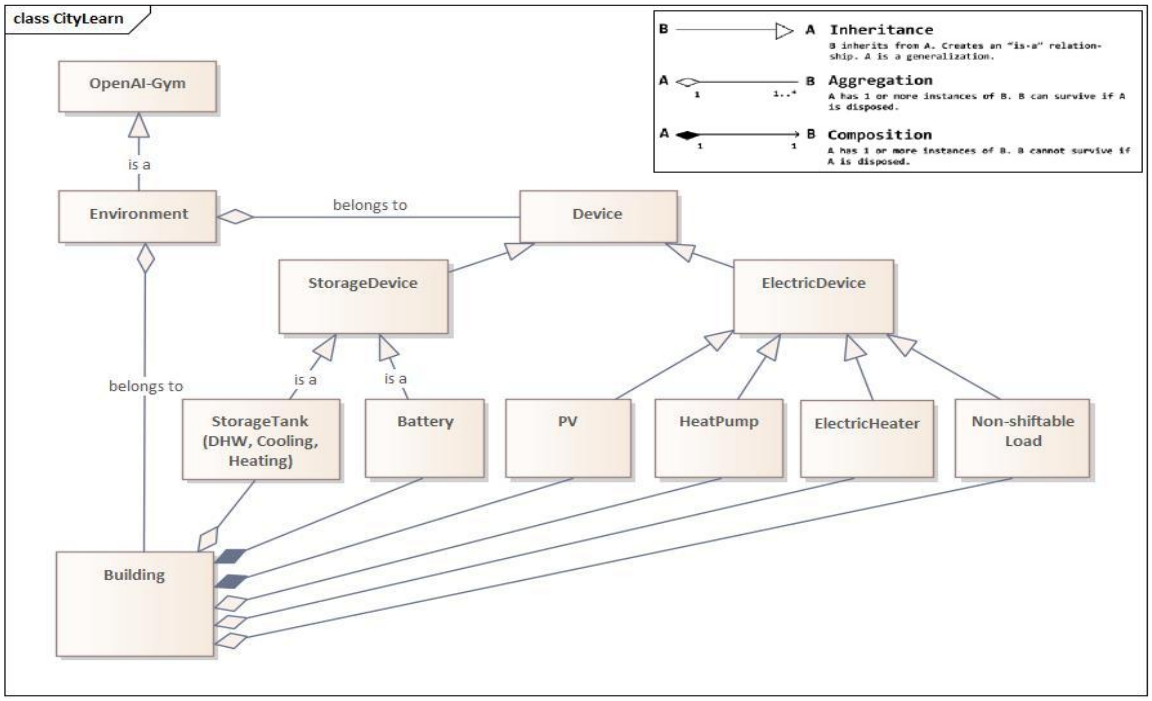

Figure 19. Component diagram of CityLearn 2.2b.

Figure 19 illustrates CityLearn's component diagram, which does not include support for EV-related domains such as EV chargers and EVs. The diagram highlights that CityLearn's original design had three main classes: *Building*, *StorageDevice*, and *ElectricDevice*.

The *Building* class includes two storage device types: the *StorageTank* child class provides the concrete implementation for components such as DHW (domestic hot water), cooling, and heating tanks; the *Battery* class implements behaviours of an active energy storage system. A storage device instance has the general attributes of an energy storage system such as energy capacity, round-trip efficiency, initial SOC, SOC, and the action to charge or discharge with a certain power value. On the other hand, a building also consists of multiple other electric devices such as a solar PV system, heat pumps for energy generation, electric heaters, and non-shiftable loads for energy consumption.

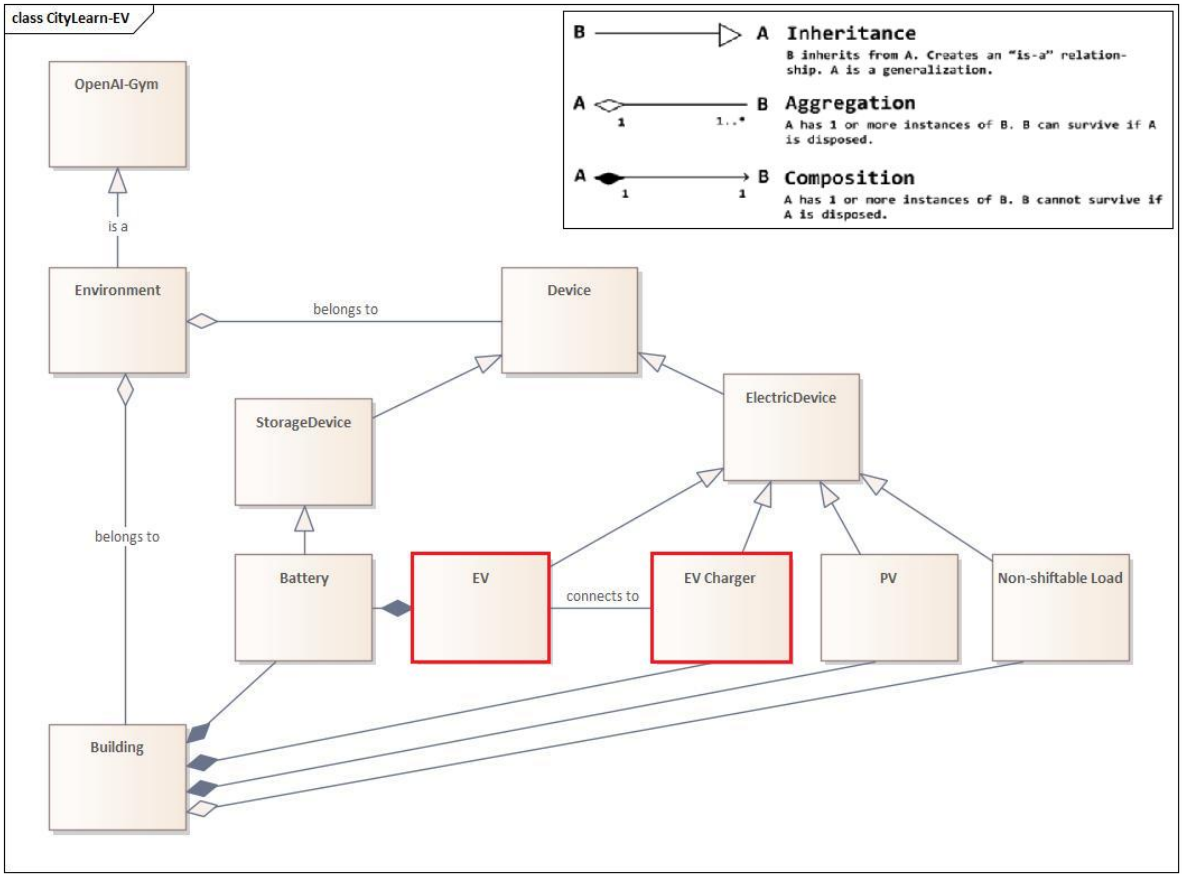

Figure 20. Implementation of EVs into CityLearn environment.

For the integration of EVs in the original CityLearn, we need to implement new models for EV-related domains i.e., the *EVCharger* and *EV* classes. The *EVCharger* class is derived from *ElectricDevice* to provide the action of EV electricity consumption in G2V mode and energy generation in V2G mode. The EV class is also a type of electric device that can function as an energy generation or energy consumption via its battery system. Figure 20 illustrates the new types of devices for EVs and their relationships to other components within the CityLearn environment.

6.2.2 RL Agent Design

CityLearn offers three possible control configurations: single agent, autonomous multi-agent, and synchronized multi-agent, as illustrated in Figure 21.

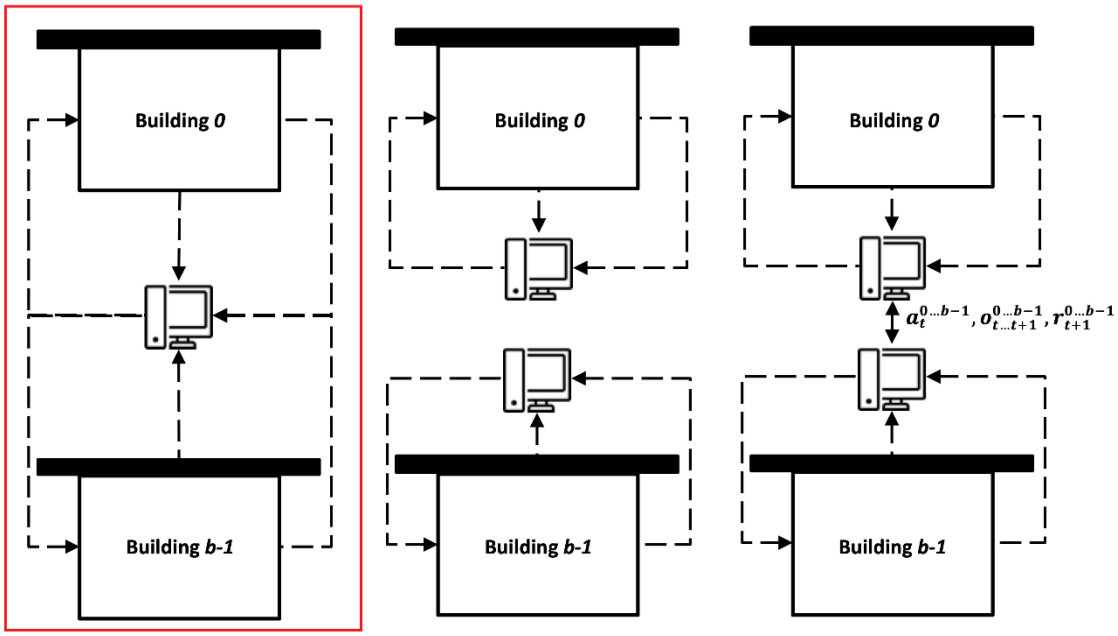

Figure 21 Control configurations (from left to right): single agent, autonomous multi-agent, and synchronized multi-agent. The study employs the single-agent control (Nweye et al., 2024).

In the single agent configuration, a centralized control agent collects observations from all buildings and prescribes actions for the district's DERs. The agent gets a reward value at each time step and learns a control policy. This setup is similar to an energy aggregator managing multiple distributed energy resources across multiple buildings. In the simulation, we employ the single agent configuration as it offers several benefits serving the study purposes:

- Centralized Control and Coordination: With a single agent overseeing all buildings and DERs in the district, there is a streamlined, centralized approach to decision-making. This allows for more effective coordination across resources, which can maximize system-wide benefits, such as overall energy efficiency and peak load reduction.
- Simplified Training Process: Since only one agent is trained, the training process is generally simpler and faster compared to multi-agent setups. The agent learns a single, unified policy that applies to the entire district, reducing the computational complexity associated with training multiple independent agents.
- Similar to Real-World Aggregators: The single-agent setup is analogous to how real-world energy aggregators manage distributed resources across a grid. This makes it a

useful configuration for simulating and testing control strategies that could be applied in practical, aggregated energy management systems.

6.2.3 RL Observation and Action Space Design

Table 5 shows that the agent's observation space comprises community-level time stamp information and weather states, as well as building-specific states. The three timestamp observations represent the time-dependent features of the environment and are common to all historical data. Weather observations include both direct sun irradiance and predicted data from the forecasting layer. Non-shiftable load is the overall building load before accounting for solar power, BESS contribution, and EV charging/discharging. Net electricity consumption is calculated as the sum of non-shiftable load, solar generation, BESS contribution, and EV chargers consumption as equation (5.1). Carbon emissions and net electricity prices capture the environmental and financial impacts of using the utility grid to meet building demands. To support the learning process, observations are transformed using cyclical transformations, min-max normalization, or one-hot encoding, depending on the data types.

Table 5. Observation space.

Observation	Unit	Transformation
DateTime		
Month	-	One-hot
Hour	-	Cyclical
Day type (weekday or weekend)	-	One-hot
Weather		
Direct solar irradiance	$\mathrm{W/m^2}$	Min-max norm.
Direct solar irradiance (+6h)	$\mathrm{W/m^2}$	Min-max norm.
Direct solar irradiance (+12h)	$\mathrm{W/m^2}$	Min-max norm.
Direct solar irradiance (+24h)	$\mathrm{W/m^2}$	Min-max norm.
Building		

Solar generation	kWh	Min-max norm.
Nonshiftable load	kWh	Min-max norm.
Batter SoC	-	Min-max norm.
net_electricity_consumption	kWh	Min-max norm.
Carbon emissions	$\mathrm{Kg}(\mathrm{CO}^2)$	Min-max norm.
Electricity price	$	Min-max norm.
Electricity price (+6h)	$	Min-max norm.
Electricity price (+12h)	$	Min-max norm.
Electricity price (+24h)	$	Min-max norm.
Electric Vehicle		
EV charger connected state	-	One-hot
Connected EV's SoC	-	Min-max norm.
Connected EV's estimated departure time	Hour	Min-max norm.
Connected EV's required SoC at departure	-	Min-max norm.
Incoming EV estimated arrival time	Hour	Min-max norm.
Incoming EV estimated SoC at arrival	-	Min-max norm.

The action space for the SAC agent is continuous and two-dimensional, with each dimension corresponding to a control action for either the stationary battery or the EV charger. One dimension of the action space controls the stationary battery, representing the fraction of the battery's capacity to be charged (positive value) or discharged (negative value). The action values for the stationary battery range from -1 to 1, where -1 means discharging at maximum power, 0 means no action, and 1 means charging at maximum power. Another dimension in the action space controls the EV charger, which also supports bi-directional power flow (both charging and discharging). Similar to the stationary battery, values for the EV charger control range between -1 and 1, where positive values indicate charging the EV, and negative values indicate discharging the EV (sending power from the EV battery back to the grid or the building). The agent must consider constraints such as maximum charge/discharge

rates for each battery, the SoC limits, and possibly a minimum charge level to avoid depleting the EV or stationary battery completely.

6.2.4 Reward Design

The main target of the SAC deep RL agent is to minimize electricity costs while ensuring the satisfaction of users regarding the required SoC of EV BESS at departure time.

$$r = \sum_{i=0}^{n} \left[\left(p_i^{battery} + p_i^{EV\_BESS} \right) \times |C_i| \right] \quad (6.1)$$

The reward function r is structured to reduce electricity costs, C, by providing feedback that accounts for all n buildings in the system. It is calculated individually for each building i and aggregated to reflect the collective performance. This function incentivizes near net-zero energy usage by penalizing situations where the grid is used to meet demand despite available battery energy and when there is excess energy export without the batteries being fully charged, through penalty terms $p_i^{battery}$ and $p_i^{EV\_BESS}$.

The former requirement regarding minimizing electricity costs by controlling the stationary battery system can be achieved by training the agent to charge the battery during times when electricity prices are low, typically between 9 PM and 4 PM when the grid also tends to have lower emissions. Each building is also capable of generating its power when solar radiation is available. By taking advantage of this self-generated solar power in the late morning and early afternoon, the battery system can be charged at no additional cost and then discharged throughout the rest of the day, resulting in lower electricity costs and emissions. Additionally, by using the batteries to cover early morning and evening peaks, we can reduce peak demand and improve the load factor, ultimately enhancing performance across the defined KPIs. We should guide the agent to make full use of solar energy by charging the batteries with PV generation whenever they have the capacity available. On the other hand, the agent should

be trained to discharge energy when there is excess demand on the grid and sufficient energy remains in the batteries.

$$p_i^{battery} = -\left(1 + sign(C_i) \times SoC_i^{battery}\right) \quad (6.2)$$

When the batteries are fully charged and exporting energy to the grid, no penalty or reward is applied. In contrast, if the battery is at full capacity and energy is imported from the grid, the penalty reaches its maximum level.

The reward function must also address the requirement regarding the required SoC of connected EVs at departure time. Penalty term is calculated for the EV charger based on how well the SoC of the EV aligns with desired conditions, specifically focusing on penalizing inefficient or unnecessary charging behaviors.

$$p_i^{EV\_BESS} = \begin{cases} -SoC_{diff} \div ev_{departure\_time} \times \omega_1 & , when\ significant\ SoC\ deficit \\ -SoC_{diff} \div ev_{departure\_time} \times \omega_2 \times sign(C_i) & , when\ moderate\ SoC\ deficit \\ -\left(1 + sign(C_i) \times SoC_i^{EV_{BESS}}\right) & , when\ SoC\ surplus \end{cases} \quad (6.3)$$

The penalty term for EV BESS reflects that If the SOC deficit SoC_{diff} exceeds the maximum possible charge within the given time, a larger penalty is applied by the coefficient ω_1. It emphasizes the SoC difference with a heavier penalty. The calculations also effectively discourage charging inefficiency when there is an SoC deficit to be reasonably met before departure by scaling the penalty by the EV departure time, the time left until the EV departs. In another case when the SoC deficit is within a manageable range, a smaller penalty is applied while also addressing the factor of whether charging is costly or not by multiplying the electricity costs sign, $sign(C_i)$. Last but not least, the penalty for SoC surplus, is calculated similarly to the penalty for the stationary battery. If charging would be beneficial based on $sign(C_i)$, a reward is granted to incentivize reducing grid demand when the SoC is sufficient.

6.2.5 Designed SAC Compared to Other Control Algorithms

In the simulation, the performance of the designed SAC agent is evaluated by comparing it to various control algorithms, including:

- **Baseline control:**
  In this setup, there is no intelligent control applied to the system. The battery and EV charger operate without any optimization, meaning they may charge or discharge in a default manner or not at all, based solely on unoptimized demand patterns. This baseline provides a reference point to measure how much energy cost and efficiency improvements the SAC agent and other control strategies offer compared to having no smart control in place.

- **Random Rule-Based Control (RBC):**
  In this control, actions are chosen at random following a basic set of rules for the battery and EV charger. Although there are guidelines for when to charge or discharge, the specific actions are not tailored to optimize cost, emissions, or performance.

- **Optimized Rule-Based Control (RBC):**
  This version of RBC applies a structured, rule-based approach with optimized heuristics. It is based on manual configuration and has no flexibility at all, hence does not have the capabilities to adapt to the changes in the EV charging schedule, and variability of the loads. In this strategy, the EV charger is designed to mainly charge EV BESS during night time and in the early morning (10 PM – 8 AM) while the discharging is configured to be carried out only during 5 PM – 10 PM when the electricity price is potentially high. This optimized RBC is typically more efficient than random RBC and serves as a benchmark for how well a manually tuned, rules-based approach performs compared to the designed SAC agent.

- **Default SAC (No Designed Reward for EV):**

A SAC agent controls the battery and EV charger without a reward structure specifically designed for EV optimization. The agent is trained to optimize general performance metrics, but it does not take into account detailed EV charging goals, such as preventing overcharging or meeting specific SoC requirements. With this design of the reward function, we can predict that it always charges or discharges all the time as there is no control tailoring for EV BESS.

6.3 Simulation Results

6.3.1 SAC Models Training Convergence

In the analysis outcome, the convergence speed, or the rate at which the agent reaches optimal performance, between the default SAC (without EV-specific rewards) and the optimized SAC (with a tailored reward structure) shows no significant difference.

Figure 22 illustrates that both versions of the SAC agent, despite having different reward designs, take a similar number of training steps or episodes to reach their peak performance. This indicates that adding specialized reward terms for EV charging and discharging does not slow down or accelerate the learning process itself. Rather, it adjusts the agent's focus within the same timeframe, helping it achieve more specific goals, such as managing EV state-of-charge more effectively or minimizing grid interaction costs.

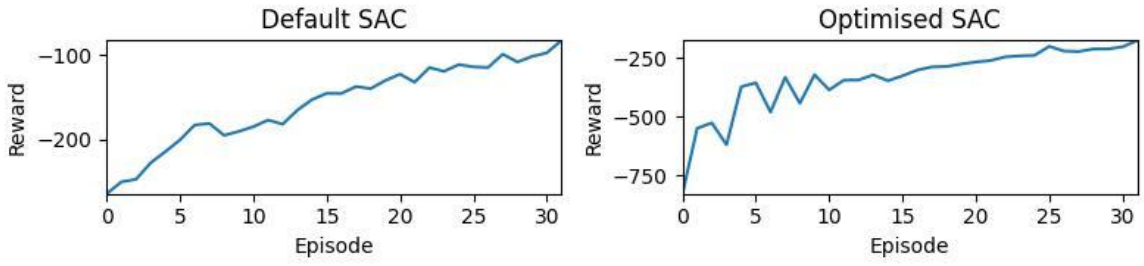

Figure 22. Cumulative rewards of Default SAC and Optimized SAC.

In essence, the tailored reward structure refines the agent's behaviour and outcomes, but it does not impact the speed at which the SAC algorithm converges. This suggests that the core SAC architecture is robust enough to handle complex, customized reward functions without sacrificing training efficiency.

6.3.2 Performance Evaluation Through Predefined KPIs

Figures 23 and 24 show that the optimized SAC agent (purple colour) demonstrates significantly improved performance over the default SAC, and other control strategies in both building-level and district-level KPIs, specifically targeting energy costs, CO_2 emissions, peak management, and load efficiency.

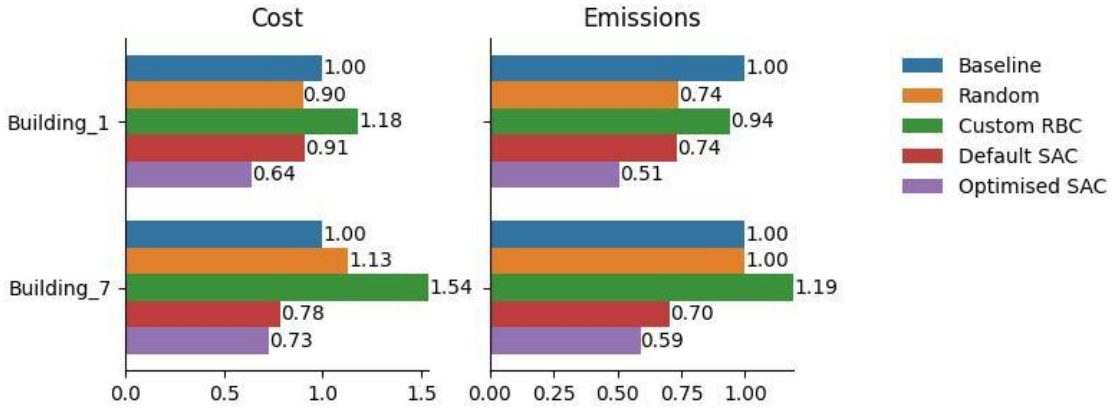

Figure 23. Building-level KPIs.

The optimized SAC agent achieves lower electricity costs at the building level by strategically managing battery storage and EV charging. Its reward structure prioritizes charging during low-cost periods and discharging during peak demand and high electricity prices, resulting in a more economical energy usage pattern for each building, hence at the district level as well. It is also noticeable in Figure 25 and Fingure 26 that the optimized SAC lowers average daily peak demand by discharging stored energy in the stationary batteries and EV BESS during peak times across buildings, especially during the second half of the day (2 PM – 11 PM). This coordinated approach helps avoid high-cost, high-emission periods and reduces peak strain on the grid.

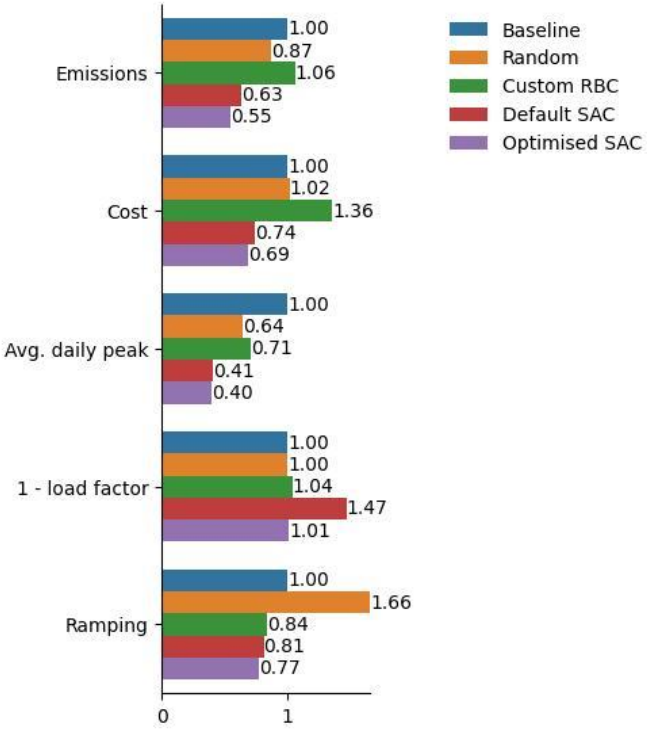

Figure 24. District-level KPIs.

Figures 25 and 26 highlight an improved load factor by smoothing demand across buildings. The optimized SAC also minimizes ramping, ensuring a steadier energy demand profile across the district. Managing the stationary batteries and EV chargers smoothly, reduces the frequency and intensity of load changes, contributing to grid stability.

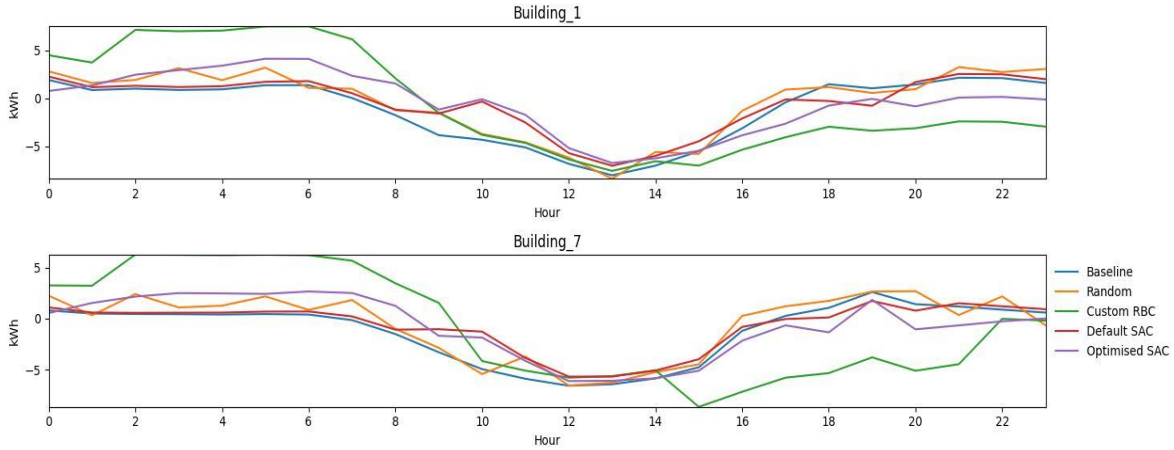

Figure 25. Building-level daily-average load profiles.

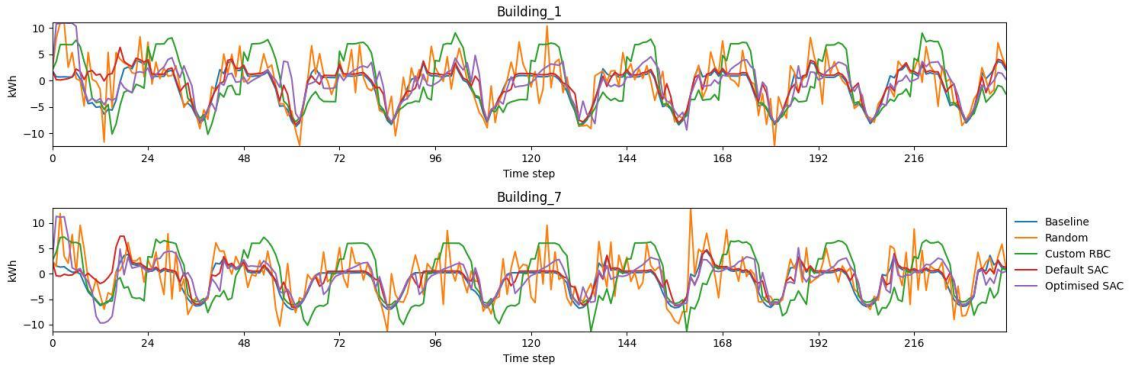

Figure 26. Building-level daily load profiles during simulation time steps.

6.3.3 Performance Evaluation with Stationary BESS and EV BESS

In Figure 27, the red points indicate the desired EV BESS charge level at the departure time. The comparison highlights whether the agent can effectively control charging/discharging the EV BESS to ensure the satisfaction of EV users.

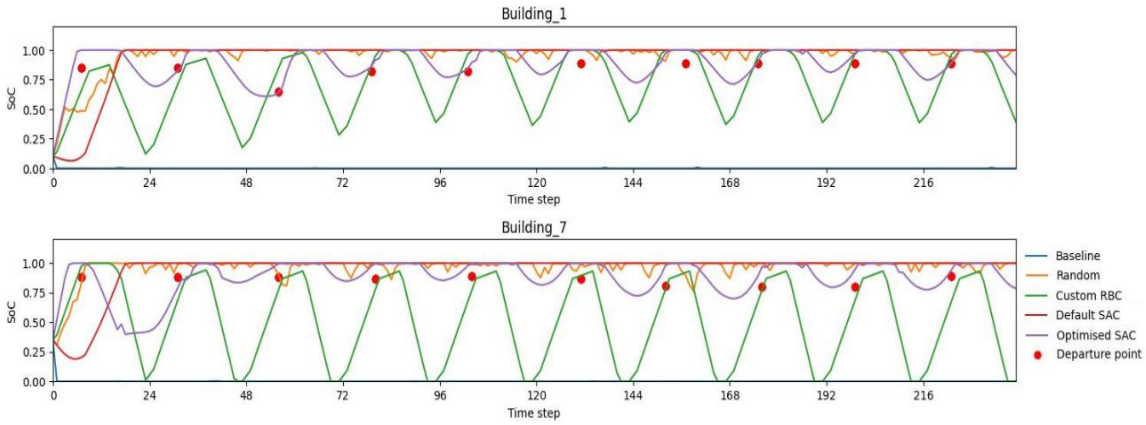

Figure 27. EV BESS SoC profiles.

The EV BESS SoC profiles illustrate that the optimized SAC provides flexible, intelligent control that not only meets building and district-level KPIs such as cost savings, emissions reduction, and load balancing but also ensures the EV is consistently charged to the desired SoC level

before users need it. This balance reflects the optimized SAC's ability to dynamically respond to demand and price fluctuations, prioritizing efficient charging while meeting user needs. The default SAC approach, lacking tailored rewards for EV SoC optimization, tends to prioritize maintaining the EV SoC at the maximum level. This behaviour means that the EV BESS is frequently fully charged, even during peak demand times, which can drive up costs and emissions due to inflexible grid reliance. The default SAC, therefore, lacks the nuanced control needed to optimize charging based on time-of-use rates and grid conditions, missing opportunities to charge more cost-effectively and sustainably.

The custom RBC approach, although designed to improve energy efficiency, misses certain demand levels for the EV SoC. Without the adaptability of RL, the RBC lacks the ability to anticipate user needs and dynamically adjust charging in real-time, resulting in suboptimal SoC levels at times when the EV is expected to be ready. This rigidity leads to unmet SoC targets, falling short of ensuring user readiness and overall efficiency.

The plots below illustrate the state of charge (SoC) profiles for stationary batteries in two buildings under different control strategies, highlighting how each approach manages battery usage over the simulation time steps. The purple line represents the optimized SAC control strategy, characterized by a highly dynamic and adaptive behaviour. The SoC fluctuates significantly, reaching both high and low states as needed, reflecting efficient energy utilization.

In contrast, the red line represents the default SAC strategy, which is less dynamic compared to the optimized SAC. Although it adjusts the SoC over time, it does not utilize the battery’s full capacity as efficiently. Additionally, during the initial simulation period, the default SAC strategy appears to overextend the battery capacity by persistently discharging, resulting in lower SoC levels.

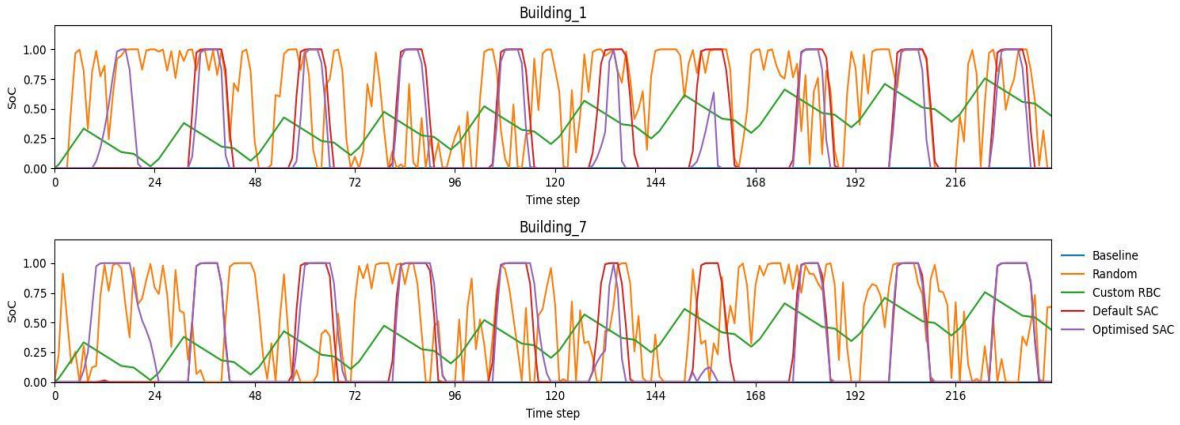

Figure 28 Stationary batteries SoC profiles.

7 Discussion and Conclusions

This research contributes to the growing field of deep RL for intelligent energy management in grid-interactive buildings. By focusing on the development of specialized RL-based energy management systems, the study demonstrates how flexible control of distributed energy resources (DERs), especially the integration of EVs, can empower prosumers to reduce energy costs, enhance grid stability, optimize energy use, and integrate renewable energy sources more effectively. The findings underscore the potential for RL-based systems to enable more efficient energy management at both the building and grid levels. Additionally, the study provides insights into the practical challenges and benefits of prosumer engagement in modern energy systems, highlighting the need for further refinement to ensure robust, real-world implementation. Overall, this research advances our understanding of how intelligent control mechanisms can drive sustainable and economically viable energy practices in grid-interactive buildings.

7.1 Application of RL in the Context of Energy Management in GEBs

By analyzing 143 key studies from the past decade on RL applications in building energy management, using bibliometric analysis, particularly co-citation techniques, the present research identified nine main research groups spanning topics from RL fundamentals, Q-Learning for discrete actions, and advanced RL methods supporting continuous actions, to more specialized applications in building energy and home energy management with a focus on HVAC control. It also covered RL and deep neural networks (DNN) integration in household energy management systems, deterministic policy gradient control for smart homes, advanced RL techniques for BEM optimization, and demand response optimization using RL. For future work, the review highlights a need to assess RL algorithms' impact on carbon emissions, noting that few studies include emissions as a performance metric. Additionally, a gap exists for studies that incorporate comprehensive energy systems, including renewable energy sources, EV chargers, diverse load types, batteries, and thermal systems, to allow for

broader comparisons of RL methods like RBC, Q-Learning, DQN, and SAC across multiple key performance indicators.

7.2 Simulation of SAC – Deep RL for Intelligent Energy Management in GEBs

The simulation analysis reveals that the optimized SAC agent outperforms other control strategies in building and district-level KPIs. It excels in reducing energy costs and CO_2 emissions, managing peaks, and improving load efficiency. Controlling stationary batteries and EV chargers efficiently, decreases the frequency and intensity of load fluctuations, supporting grid stability.

The optimized SAC also achieves lower electricity costs by charging during low-cost periods and discharging during peak demand times, resulting in a cost-effective usage pattern for buildings and the district overall. By discharging stored energy in the late afternoon and evening, it minimizes average daily peaks, reduces grid strain, and aligns with high-cost periods.

In contrast, the default SAC often maintains EV SoC at maximum, driving up costs due to inflexible grid reliance. The custom RBC, lacking RL's adaptability, fails to meet optimal SoC levels at times, limiting readiness and efficiency.

7.3 Challenges of Implementing RL in a Real-World Environment

Implementing RL in real-world environments poses several significant challenges. While the proposed framework of SAC single-agent RL and simulation results has shown promise, transitioning these methods to practical, real-world applications involves unique complexities.
The first major challenge is regarding the data quality and availability. RL agents require vast amounts of high-quality, relevant data to learn effective policies, which can be difficult to obtain in real-world environments. For instance, in energy management systems, detailed data on energy usage, grid status, and weather patterns are needed, but data quality may

vary, and historical records may be incomplete (Wang and Hong, 2020b). This challenge raises questions about data collection practices, the need for standardized data formats, and the potential for using simulated environments to supplement real-world data. Additionally, designing agents that can handle unformatted live data becomes essential in overcoming this hurdle. In this study, the proposed optimized SAC agent shows potential performance using one-year data including simulated data for EVs schedule and predicted data for solar irradiance and electricity price.

The next problem that needs to be addressed is the ability to generalization to dynamic real-world environments. Real-world environments are often dynamic and unpredictable. For instance, energy demand patterns, weather conditions, electricity pricing, and user behavior can vary widely, making it hard for RL agents trained in static or simulated environments to generalize well. Adaptive RL techniques, such as meta-learning or transfer learning, could help address this by enabling agents to adjust to new scenarios (Wang and Hong, 2020b). However, finding a balance between generalization and specialization remains a technical and computational challenge.

References

Afram, A., Janabi-Sharifi, F., 2014. Theory and applications of HVAC control systems – A review of model predictive control (MPC). Build. Environ. 72, 343–355. https://doi.org/10.1016/j.buildenv.2013.11.016

Afroz, Z., Shafiullah, G., Urmee, T., Higgins, G., 2018. Modeling techniques used in building HVAC control systems: A review. Renew. Sustain. Energy Rev. 83, 64–84. https://doi.org/10.1016/j.rser.2017.10.044

Ali, M., Prakash, K., Hossain, M.A., Pota, H.R., 2021. Intelligent energy management: Evolving developments, current challenges, and research directions for sustainable future. J. Clean. Prod. 314, 127904. https://doi.org/10.1016/j.jclepro.2021.127904

Bayasgalan, A., Park, Y.S., Koh, S.B., Son, S.-Y., 2024. Comprehensive Review of Building Energy Management Models: Grid-Interactive Efficient Building Perspective. Energies 17, 4794. https://doi.org/10.3390/en17194794

Bellman, R., 2010. Dynamic Programming, Princeton Landmarks in Mathematics and Physics. Princeton: Princeton University Press.

Brandi, S., Gallo, A., Capozzoli, A., 2022. A predictive and adaptive control strategy to optimize the management of integrated energy systems in buildings. Energy Rep. 8, 1550–1567. https://doi.org/10.1016/j.egyr.2021.12.058

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W., 2016. OpenAI Gym.

Dean, J., Voss, P., Gagne, D., Vasquez, D., Langner, R., 2021. Blueprint for Integrating Grid-Interactive Efficient Building (GEB) Technologies into U.S. General Services Administration Performance Contracts (No. NREL/TP-7A40-78190, 1784273, MainId:32099). https://doi.org/10.2172/1784273

Deltetto, D., Coraci, D., Pinto, G., Piscitelli, M.S., Capozzoli, A., 2021. Exploring the Potentialities of Deep Reinforcement Learning for Incentive-Based Demand Response in a Cluster of Small Commercial Buildings. Energies 14, 2933. https://doi.org/10.3390/en14102933

Deng, X., Zhang, Y., Zhang, Y., Qi, H., 2022. Toward Smart Multizone HVAC Control by Combining Context-Aware System and Deep Reinforcement Learning. IEEE Internet Things J. 9, 21010–21024. https://doi.org/10.1109/JIOT.2022.3175728

Drgoňa, J., Arroyo, J., Cupeiro Figueroa, I., Blum, D., Arendt, K., Kim, D., Ollé, E.P., Oravec, J., Wetter, M., Vrabie, D.L., Helsen, L., 2020. All you need to know about model predictive control for buildings. Annu. Rev. Control 50, 190–232. https://doi.org/10.1016/j.arcontrol.2020.09.001

Du, Y., Zandi, H., Kotevska, O., Kurte, K., Munk, J., Amasyali, K., Mckee, E., Li, F., 2021. Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. Appl. Energy 281, 116117. https://doi.org/10.1016/j.apenergy.2020.116117

Fliess, M., Join, C., 2013. Model-free control. Int. J. Control 86, 2228–2252. https://doi.org/10.1080/00207179.2013.810345

Fontenot, H., Dong, B., 2019. Modeling and control of building-integrated microgrids for optimal energy management – A review. Appl. Energy 254, 113689. https://doi.org/10.1016/j.apenergy.2019.113689

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., Levine, S., 2019. Soft Actor-Critic Algorithms and Applications.

Hossain, E., Khan, I., Un-Noor, F., Sikander, S.S., Sunny, Md.S.H., 2019. Application of Big Data and Machine Learning in Smart Grid, and Associated Security Concerns: A Review. IEEE Access 7, 13960–13988. https://doi.org/10.1109/ACCESS.2019.2894819

Ipakchi, A., Albuyeh, F., 2009. Grid of the future. IEEE Power Energy Mag. 7, 52–62. https://doi.org/10.1109/MPE.2008.931384

Jensen, S.Ø., Marszal-Pomianowska, A., Lollini, R., Pasut, W., Knotzer, A., Engelmann, P., Stafford, A., Reynders, G., 2017. IEA EBC Annex 67 Energy Flexible Buildings. Energy Build. 155, 25–34. https://doi.org/10.1016/j.enbuild.2017.08.044

Jo, T., 2021. Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-65900-4

Jung, A., 2022. Machine Learning: The Basics, Machine Learning: Foundations, Methodologies, and Applications. Springer Nature Singapore, Singapore. https://doi.org/10.1007/978-981-16-8193-6

Li, B., Roche, R., Miraoui, A., 2017. Microgrid sizing with combined evolutionary algorithm and MILP unit commitment. Appl. Energy 188, 547–562. https://doi.org/10.1016/j.apenergy.2016.12.038

Li, H., Wan, Z., He, H., 2020. Real-Time Residential Demand Response. IEEE Trans. Smart Grid 11, 4144–4154. https://doi.org/10.1109/TSG.2020.2978061

Li, H., Wang, Z., Hong, T., Piette, M.A., 2021. Energy flexibility of residential buildings: A systematic review of characterization and quantification methods and applications. Adv. Appl. Energy 3, 100054. https://doi.org/10.1016/j.adapen.2021.100054

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2019. Continuous control with deep reinforcement learning.

Lu, R., Hong, S.H., 2019. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. Appl. Energy 236, 937–949. https://doi.org/10.1016/j.apenergy.2018.12.061

Lu, R., Hong, S.H., Yu, M., 2019. Demand Response for Home Energy Management Using Reinforcement Learning and Artificial Neural Network. IEEE Trans. Smart Grid 10, 6629–6639. https://doi.org/10.1109/TSG.2019.2909266

Lu, R., Hong, S.H., Zhang, X., 2018. A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. Appl. Energy 220, 220–230. https://doi.org/10.1016/j.apenergy.2018.03.072

Lu, R., Jiang, Z., Wu, H., Ding, Y., Wang, D., Zhang, H.-T., 2023. Reward Shaping-Based Actor–Critic Deep Reinforcement Learning for Residential Energy Management. IEEE Trans. Ind. Inform. 19, 2662–2673. https://doi.org/10.1109/TII.2022.3183802

Macieira, P., Gomes, L., Vale, Z., 2021. Energy Management Model for HVAC Control Supported by Reinforcement Learning. Energies 14, 8210. https://doi.org/10.3390/en14248210

Mason, K., Grijalva, S., 2019. A review of reinforcement learning for autonomous building energy management. Comput. Electr. Eng. 78, 300–312. https://doi.org/10.1016/j.compeleceng.2019.07.019

Michailidis, P., Michailidis, I., Gkelios, S., Kosmatopoulos, E., 2024. Artificial Neural Network Applications for Energy Management in Buildings: Current Trends and Future Directions. Energies 17, 570. https://doi.org/10.3390/en17030570

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533. https://doi.org/10.1038/nature14236

Mocanu, E., Mocanu, D.C., Nguyen, P.H., Liotta, A., Webber, M.E., Gibescu, M., Slootweg, J.G., 2019. On-Line Building Energy Optimization Using Deep Reinforcement Learning. IEEE Trans. Smart Grid 10, 3698–3708. https://doi.org/10.1109/TSG.2018.2834219

Mwasilu, F., Justo, J.J., Kim, E.-K., Do, T.D., Jung, J.-W., 2014. Electric vehicles and smart grid interaction: A review on vehicle to grid and renewable energy sources integration. Renew. Sustain. Energy Rev. 34, 501–516. https://doi.org/10.1016/j.rser.2014.03.031

Nagy, Z., Park, J.Y., Vázquez-Canteli, J.R., 2018. Reinforcement learning for intelligent environments: A tutorial, in: Routledge Handbook of Sustainable and Resilient Infrastructure. Routledge.

Nakabi, T.A., Toivanen, P., 2021. Deep reinforcement learning for energy management in a microgrid with flexible demand. Sustain. Energy Grids Netw. 25, 100413. https://doi.org/10.1016/j.segan.2020.100413

Neukomm, M., Nubbe, V., Fares, R., 2019. Grid-Interactive Efficient Buildings Technical Report Series: Overview of Research Challenges and Gaps (No. NREL/TP--5500-75470, DOE/GO--102019-5227, 1577966). https://doi.org/10.2172/1577966

Nweye, K., Kaspar, K., Buscemi, G., Fonseca, T., Pinto, G., Ghose, D., Duddukuru, S., Pratapa, P., Li, H., Mohammadi, J., Ferreira, L.L., Hong, T., Ouf, M., Capozzoli, A., Nagy, Z., 2024. CityLearn v2: Energy-flexible, resilient, occupant-centric, and carbon-aware management of grid-interactive communities.

Ojand, K., Dagdougui, H., 2022. Q-Learning-Based Model Predictive Control for Energy Management in Residential Aggregator. IEEE Trans. Autom. Sci. Eng. 19, 70–81. https://doi.org/10.1109/TASE.2021.3091334

OpenAI, 2020. Soft Actor-Critic — Spinning Up documentation [WWW Document]. OpenAI Spinn. Up. URL https://spinningup.openai.com/en/latest/algorithms/sac.html (accessed 10.21.24).

Parhizi, S., Lotfi, H., Khodaei, A., Bahramirad, S., 2015. State of the Art in Research on Microgrids: A Review. IEEE Access 3, 890–925. https://doi.org/10.1109/ACCESS.2015.2443119

Parvin, K., Lipu, M.S.H., Hannan, M.A., Abdullah, M.A., Jern, K.P., Begum, R.A., Mansur, M., Muttaqi, K.M., Mahlia, T.M.I., Dong, Z.Y., 2021. Intelligent Controllers and Optimization Algorithms for Building Energy Management Towards Achieving Sustainable Development: Challenges and Prospects. IEEE Access 9, 41577–41602. https://doi.org/10.1109/ACCESS.2021.3065087

Pinto, G., Deltetto, D., Capozzoli, A., 2021a. Data-driven district energy management with surrogate models and deep reinforcement learning. Appl. Energy 304, 117642. https://doi.org/10.1016/j.apenergy.2021.117642

Pinto, G., Piscitelli, M.S., Vázquez-Canteli, J.R., Nagy, Z., Capozzoli, A., 2021b. Coordinated energy management for a cluster of buildings through deep reinforcement learning. Energy 229, 120725. https://doi.org/10.1016/j.energy.2021.120725

Reiner, L., 2024. rl-institut/simbev.

Ruelens, F., Claessens, B.J., Vandael, S., De Schutter, B., Babuška, R., Belmans, R., 2017. Residential Demand Response of Thermostatically Controlled Loads Using Batch Reinforcement Learning. IEEE Trans. Smart Grid 8, 2149–2159. https://doi.org/10.1109/TSG.2016.2517211

Shaikh, P.H., Nor, N.B.M., Nallagownden, P., Elamvazuthi, I., Ibrahim, T., 2014. A review on optimized control systems for building energy and comfort management of smart sustainable buildings. Renew. Sustain. Energy Rev. 34, 409–429. https://doi.org/10.1016/j.rser.2014.03.027

Shakeri, M., Shayestegan, M., Abunima, H., Reza, S.M.S., Akhtaruzzaman, M., Alamoud, A.R.M., Sopian, K., Amin, N., 2017. An intelligent system architecture in home energy management systems (HEMS) for efficient demand response in smart grid. Energy Build. 138, 154–164. https://doi.org/10.1016/j.enbuild.2016.12.026

Shen, R., Zhong, S., Wen, X., An, Q., Zheng, R., Li, Y., Zhao, J., 2022. Multi-agent deep reinforcement learning optimization framework for building energy system with renewable energy. Appl. Energy 312, 118724. https://doi.org/10.1016/j.apenergy.2022.118724

Small, H., 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. J. Am. Soc. Inf. Sci. 24, 265–269. https://doi.org/10.1002/asi.4630240406

Southern California Edison, 2022. Time-Of-Use Residential Rate Plans | Rates | Your Home | Home - SCE [WWW Document]. URL https://www.sce.com/residential/rates/Time-Of-Use-Residential-Rate-Plans

Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning, second edition: An Introduction. MIT Press.

Tang, H., Wang, S., 2021. Energy flexibility quantification of grid-responsive buildings: Energy flexibility index and assessment of their effectiveness for applications. Energy 221, 119756. https://doi.org/10.1016/j.energy.2021.119756

United Nations, 2015. THE 17 GOALS | Sustainable Development [WWW Document]. URL https://sdgs.un.org/goals (accessed 11.10.24).

Van Hasselt, H., Guez, A., Silver, D., 2016. Deep Reinforcement Learning with Double Q-Learning. Proc. AAAI Conf. Artif. Intell. 30. https://doi.org/10.1609/aaai.v30i1.10295

Vazquez-Canteli, J.R., Dey, S., Henze, G., Nagy, Z., 2020. CityLearn: Standardizing Research in Multi-Agent Reinforcement Learning for Demand Response and Urban Energy Management.

Vázquez-Canteli, J.R., Kämpf, J., Henze, G., Nagy, Z., 2019. CityLearn v1.0: An OpenAI Gym Environment for Demand Response with Deep Reinforcement Learning, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient

Buildings, Cities, and Transportation, BuildSys '19. Association for Computing Machinery, New York, NY, USA, pp. 356–357. https://doi.org/10.1145/3360322.3360998

Vázquez-Canteli, J.R., Nagy, Z., 2019. Reinforcement learning for demand response: A review of algorithms and modeling techniques. Appl. Energy 235, 1072–1089. https://doi.org/10.1016/j.apenergy.2018.11.002

Vineetha, C.P., Babu, C.A., 2014. Smart grid challenges, issues and solutions, in: 2014 International Conference on Intelligent Green Building and Smart Grid (IGBSG). Presented at the 2014 International Conference on Intelligent Green Building and Smart Grid (IGBSG), pp. 1–4. https://doi.org/10.1109/IGBSG.2014.6835208

Wang, Z., Hong, T., 2020a. Reinforcement learning for building controls: The opportunities and challenges. Appl. Energy 269, 115036. https://doi.org/10.1016/j.apenergy.2020.115036

Wang, Z., Hong, T., 2020b. Reinforcement learning for building controls: The opportunities and challenges. Appl. Energy 269, 115036. https://doi.org/10.1016/j.apenergy.2020.115036

Watkins, C.J.C.H., Dayan, P., 1992. Q-learning. Mach. Learn. 8, 279–292. https://doi.org/10.1007/BF00992698

Wei, T., Wang, Y., Zhu, Q., 2017. Deep reinforcement learning for building HVAC control, in: 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC). Presented at the 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC), pp. 1–6. https://doi.org/10.1145/3061639.3062224

Xie, J., Ajagekar, A., You, F., 2023. Multi-Agent attention-based deep reinforcement learning for demand response in grid-responsive buildings. Appl. Energy 342, 121162. https://doi.org/10.1016/j.apenergy.2023.121162

Xu, X., Jia, Y., Xu, Y., Xu, Z., Chai, S., Lai, C.S., 2020. A Multi-Agent Reinforcement Learning-Based Data-Driven Method for Home Energy Management. IEEE Trans. Smart Grid 11, 3201–3211. https://doi.org/10.1109/TSG.2020.2971427

Ye, Y., Qiu, D., Wu, X., Strbac, G., Ward, J., 2020. Model-Free Real-Time Autonomous Control for a Residential Multi-Energy System Using Deep Reinforcement Learning. IEEE Trans. Smart Grid 11, 3068–3082. https://doi.org/10.1109/TSG.2020.2976771

Yu, L., Jiang, T., Zou, Y., 2019. Online Energy Management for a Sustainable Smart Home With an HVAC Load and Random Occupancy. IEEE Trans. Smart Grid 10, 1646–1659. https://doi.org/10.1109/TSG.2017.2775209

Yu, L., Qin, S., Zhang, M., Shen, C., Jiang, T., Guan, X., 2021a. A Review of Deep Reinforcement Learning for Smart Building Energy Management. IEEE Internet Things J. 8, 12046–12063. https://doi.org/10.1109/JIOT.2021.3078462

Yu, L., Sun, Y., Xu, Z., Shen, C., Yue, D., Jiang, T., Guan, X., 2021b. Multi-Agent Deep Reinforcement Learning for HVAC Control in Commercial Buildings. IEEE Trans. Smart Grid 12, 407–419. https://doi.org/10.1109/TSG.2020.3011739

Yu, L., Xie, W., Xie, D., Zou, Y., Zhang, D., Sun, Z., Zhang, L., Zhang, Y., Jiang, T., 2020. Deep Reinforcement Learning for Smart Home Energy Management. Model-Free Real-Time Auton. Control Resid. Multi-Energy Syst. Using Deep Reinf. Learn. 7, 2751–2762. https://doi.org/10.1109/JIOT.2019.2957289

Zafar, R., Mahmood, A., Razzaq, S., Ali, W., Naeem, U., Shehzad, K., 2018. Prosumer based energy management and sharing in smart grid. Renew. Sustain. ENERGY Rev. 82, 1675–1684. https://doi.org/10.1016/j.rser.2017.07.018

Zenginis, I., Vardakas, J., Koltsaklis, N.E., Verikoukis, C., 2022. Smart Home's Energy Management Through a Clustering-Based Reinforcement Learning Approach. IEEE Internet Things J. 9, 16363–16371. https://doi.org/10.1109/JIOT.2022.3152586

Zhang, D., Li, S., Sun, M., O'Neill, Z., 2016. An Optimal and Learning-Based Demand Response and Home Energy Management System. IEEE Trans. Smart Grid 7, 1790–1801. https://doi.org/10.1109/TSG.2016.2552169

Appendices

Appendix 1. Energy Management in Grid-Interactive Efficient Buildings.

GEBs are interconnected, enabling two-way communication between the building and the grid, allowing signals to be exchanged that either directly manage equipment or provide information about pricing and grid conditions. These signals help trigger automated systems within the building to respond based on cost efficiency and user preferences. GEBs also incorporate intelligent systems, using sensors, controls, and data analytics to optimize building operations in alignment with occupant needs while providing valuable services to the grid. Moreover, GEBs are designed to be highly flexible, and capable of adjusting loads or tapping into DERs rapidly to ensure optimal performance and adaptability.

GEBs are equipped with advanced technologies, including sensors, automation systems, and AI-driven energy management algorithms that enable real-time monitoring and control of energy use (Bayasgalan et al., 2024). These buildings are designed to dynamically respond to signals from the electric grid, such as demand response requests, changes in electricity prices, and fluctuations in renewable energy generation. The goal of GEBs is to make buildings active participants in the energy system, capable of not only consuming energy but also producing and storing it.

At the core of GEBs is the use of intelligent energy management systems (iEMS) that optimize energy usage by integrating data from multiple sources, including weather forecasts, occupancy patterns, and grid conditions. These systems ensure that energy is used efficiently, that renewable energy is prioritized when available, and energy demand is shifted during peak hours to reduce strain on the grid (Jensen et al., 2017).

Renewable Energy Sources in GEBs

The seamless integration of renewable energy sources such as solar photovoltaic (PV) systems is one of the most important components of energy management in GEBs (Neukomm

et al., 2019). As buildings increasingly adopt onsite renewable energy generation, they can function as prosumers—both producers and consumers of electricity (Zafar et al., 2018). This enables GEBs to reduce their dependence on the grid and to contribute excess renewable energy back into the grid, enhancing grid stability.

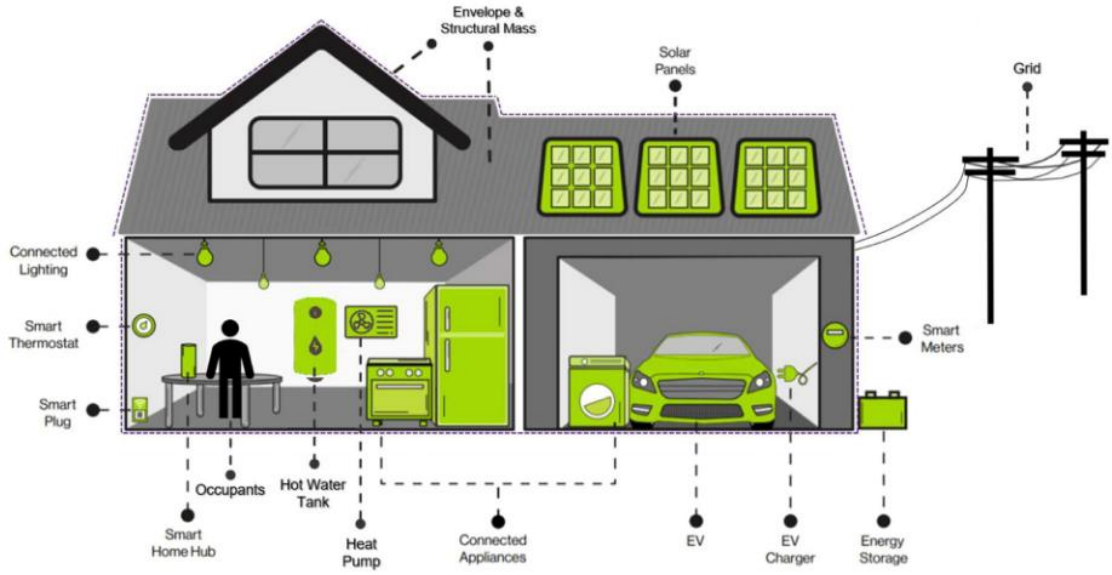

Figure 29. Typical components of a GEB (Li et al., 2021).

Solar PV systems are the most commonly integrated renewable energy technology in GEBs, especially in buildings with large roof spaces (Dean et al., 2021). Solar panels generate electricity during daylight hours, which can be used to meet the building's energy needs or stored in battery storage systems for later use, especially during the evening when solar energy is unavailable. GEBs equipped with energy storage systems can store excess energy and discharge it during peak demand periods, reducing grid congestion and avoiding the need for expensive peak-time electricity from fossil-fuel-based power plants.

Moreover, wind energy, although more site-specific, can also be integrated into GEBs in regions with favourable wind conditions (Bayasgalan et al., 2024). The inclusion of these renewable sources reduces buildings' carbon footprint while enhancing energy resilience, allowing GEBs to maintain operations even during power outages.

Electric Vehicles in GEBs

Integrating EVs and vehicle-to-grid (V2G) technology is another essential component of GEBs (Neukomm et al., 2019). EVs, which are becoming more and more popular as part of the worldwide movement toward clean transportation, are viewed as both mobile energy storage devices and a means of transportation. When demand is low or power rates are low, the large quantity of energy that EV batteries can store can be used for charging; when demand is high, the energy can be released back into the grid or the building (Mwasilu et al., 2014).

The bi-directional energy flow between GEBs and EVs through V2G technology is crucial in balancing energy demand and response. When integrated into the building's EMS, EVs can act as distributed energy resources (DERs) (Mwasilu et al., 2014). For instance, during peak hours or grid stress events, EVs can discharge stored energy back into the building to reduce their reliance on grid electricity, helping to flatten demand peaks and enhance grid stability. Similarly, when renewable energy generation is high, EVs can be charged using excess solar power, further promoting the use of clean energy.

GEBs Energy Flexibility

The ability of a building to dynamically modify its energy generation, consumption, and storage in response to a variety of internal and external factors is known as energy flexibility in GEBs. Because of this flexibility, GEBs can support the electrical grid while optimizing their energy use, guaranteeing grid stability, cost savings, and efficiency without sacrificing occupant comfort (Jensen et al., 2017). Thanks to DERs such as solar PV generation, batteries, and EVs, energy flexibility can be achieved through several mechanisms: load shifting, load shedding, moderate regulation, and intelligent energy management controls (Li et al., 2021). During times of high demand, GEBs temporarily lower their electricity use through load shedding. Contrarily, load shifting entails purposefully adjusting the timing of electricity consumption in order to reduce grid demand during normal peak hours. This strategy enables GEBs to make use of cleaner and less expensive energy sources. Furthermore, load shedding can occasionally result in the reduced load being redistributed to other times. The term "moderate

regulation" describes load regulation that happens at a minute-by-minute pace. It is a useful tool for applications in current power pricing systems, including 5-minute or 15-minute real-time pricing (RTP), which usually settles within brief periods (Tang and Wang, 2021).

The figure below illustrates the evolution of a building's profile as it improves energy efficiency, integrates on-site generation, and offers load shedding and shifting services to the grid.

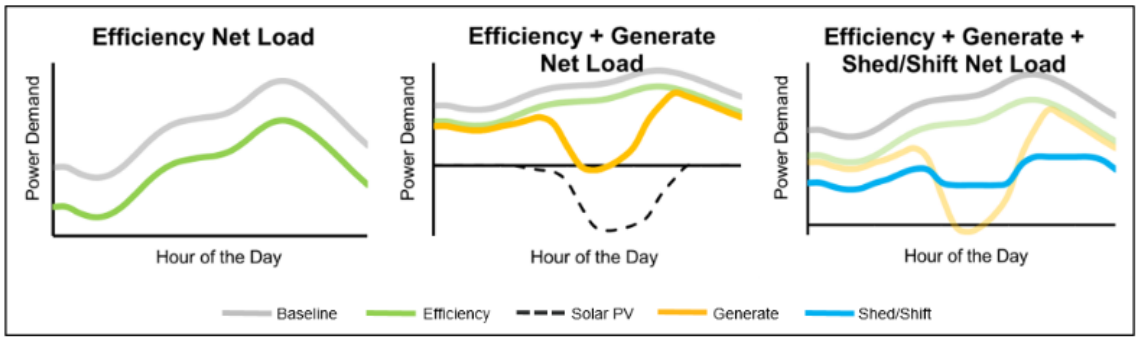

Figure 30. GEB load curves (Neukomm et al., 2019).

Specifically, load shifting can be accomplished by managing active storage systems in conjunction with renewable energy generation. Intelligent control systems can adjust to shifts in occupant behaviour, weather variations, or external signals—such as adjusting temperature set points for demand response—while maintaining energy efficiency and ensuring occupant comfort (Nweye et al., 2024).

Challenges for Energy Management in GEBs

Several challenges hinder the effective implementation and scalability of energy management systems in GEBs. These challenges arise from technical, economic, regulatory, and behavioural factors, each of which presents significant obstacles to the seamless integration of GEBs into smart energy grids. Some of the major challenges could be:

- Real-time decision-making in a complex and dynamic environment: A GEB that integrates RES like solar PV and EVs operates in a highly dynamic environment due to the

intermittent nature of RES and the unpredictable EVs' arrival and departure schedules (Ali et al., 2021). In addition to this challenge, the real-time electricity market often experiences price fluctuations due to grid conditions. An intelligent energy management system must respond to these price signals and generate appropriate control actions such as storing energy, shifting loads, or supplying excess electricity to the grid. Managing these decisions in real time is critical for minimizing costs while ensuring energy supply.

- Limitations of conventional control approaches: Traditional control strategies like Model-Based Predictive Control (MPC) and Rule-Based Control (RBC) face significant challenges when applied to complex environments, particularly in scenarios with high uncertainty and randomness in key parameters (Yu et al., 2020). These parameters include renewable energy generation, the power demand of non-shiftable loads, electricity prices, weather conditions, and the schedules of EVs. Additionally, the nature of these uncertainties, combined with the need for continuous control actions and large state spaces, further complicates the use of traditional methods. While conventional methods like RBC are inadequate for managing these uncertainties due to their inability to model or respond to continuous changes, more sophisticated approaches such as MPC or even Q-Learning RL, struggle when the action and state spaces are too large, requiring higher solver complexity and longer computation times, which can make real-time decision-making unfeasible.

- Lack of Real-Time Data Availability: Effective energy management in GEBs requires access to real-time data, such as occupancy patterns, EVs’ usage patterns weather conditions, and energy prices (Parvin et al., 2021). However, delays in data collection and processing (latency) can impact the system’s ability to respond promptly to dynamic conditions, reducing its efficiency.

- Balancing between users' comfort and energy efficiency: The contradiction between comfort and energy goals makes this difficult for the EMS because comfort increases are directly linked to rising electricity prices and consumption (Shaikh et al., 2014).

Appendix 2. Energy Management Approach for GEBs

Effective building energy management is essential for optimizing energy use, ensuring occupant comfort, and reducing operational costs. Various control strategies are employed to manage energy systems in buildings, each with its strengths, limitations, and applications. The main control approaches for building energy management include rule-based control (RBC), model-based predictive control (MPC), and model-free control. Each approach is designed to regulate system components such as HVAC, lighting, appliances, energy storage systems, and EVs, and their effectiveness depends on the complexity of the building, the data available, and the desired outcomes.

Rule-Based Control (RBC)

Rule-based control (RBC) is one of the simplest and most widely used strategies in building energy management. This approach relies on predefined rules or IF-THEN-ELSE logic to control building systems based on sensor data and operating conditions. For instance, if the electricity spot price is less than 0.001 cents and the battery SOC is low, use electricity from the grid and charge the battery by 90% of its capacity.

Despite the common adoption of this method, particularly in commercial buildings, EMS utilizing RBC cannot make adaptive decisions (Fontenot and Dong, 2019). In some cases, a RBC strategy has been found to lead to higher costs compared to other approaches (Li et al., 2017). In general, RBC strategies are better than the baseline scenario where the system has no control at all, but they fall short compared to more advanced approaches (Fontenot and Dong, 2019). This is because RBC lacks the flexibility to make intelligent decisions in dynamic environments. However, its simplicity makes it easy to implement. As demonstrated by Shakeri et al., 2017, RBC can work well for a home energy management system (HEMS) in a single residence with minimal DERs, such as a solar PV system and a BESS. For more complex GEBs including multiple DERs, especially the integration of EVs, RBC struggles to match the performance of more advanced control strategies.

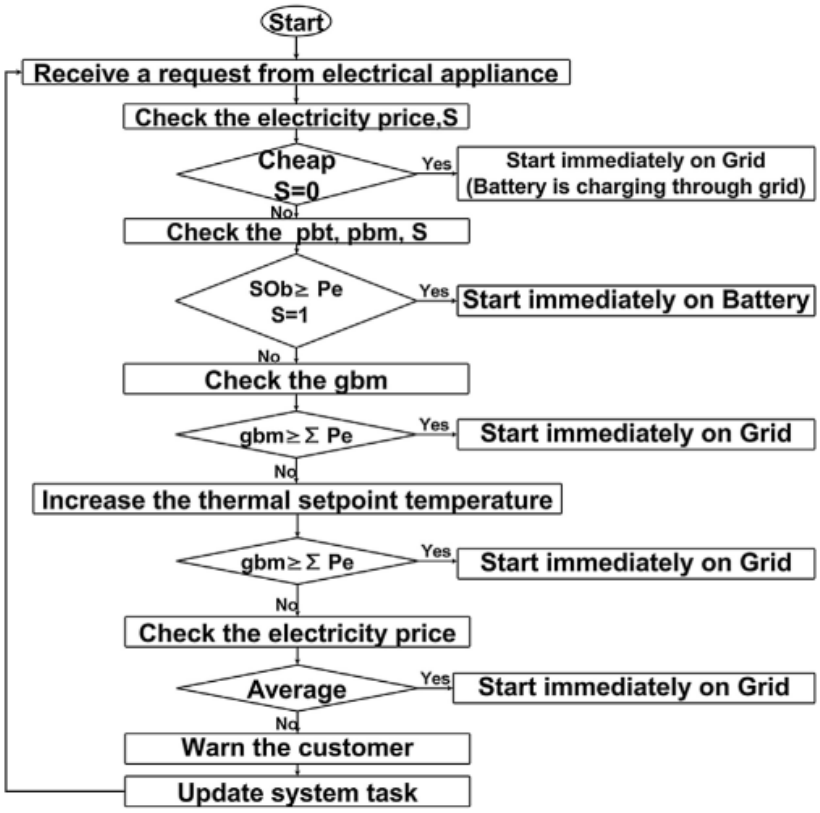

Figure 31. An RBC in HEMS (Shakeri et al., 2017).

Model-Based Predictive Control (MPC)

MPC is a sophisticated control approach that makes use of simulation tools and mathematical models of the building's energy system (such as the HVAC, lighting, etc.) to estimate how the system will act over time in response to specific inputs like weather forecasts, occupancy statistics, or energy pricing (Drgoňa et al., 2020). Based on these predictions, MPC optimizes control decisions to minimize energy use and GHG emissions while maintaining occupant comfort. It considers multiple objectives, such as reducing energy costs or minimizing carbon emissions, and adjusts system operations accordingly. MPC is widely used in complex and dynamic environments, such as large commercial buildings, where multiple variables (e.g., weather, occupancy, energy prices) affect energy use. It is often used in controlling HVAC systems, as it can optimize temperature settings and energy use while accounting for external conditions and future needs (Afram and Janabi-Sharifi, 2014).

Figure 32 shows a common abstract closed-loop MPC scheme, which can be applied to most building control systems. The control loop includes the building, influenced by disturbances d (such as weather conditions), which are forecasted by weather forecasts $\hat{d}$. The state estimator generates the state estimates $\hat{x}$, while the MPC controller optimally fine-tunes control actions u to minimize energy consumption and maintain the output vector y (e.g., indoor temperatures) within predefined comfort limits (Drgoňa et al., 2020).

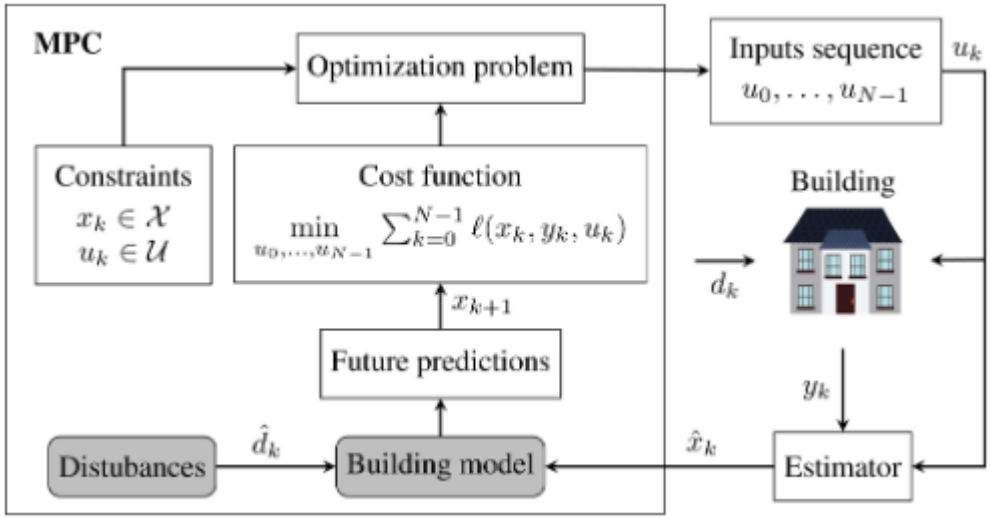

Figure 32. Diagrammatic illustration of the typical closed-loop building controller using a state estimator and MPC (Drgoňa et al., 2020).

MPC has been proven to be one of the most effective and accurate control approaches for BEMS in the existing research due to its ability to shift loads away from peak hours, adapt to unforeseen disruptions, and efficiently utilize the thermal mass of buildings by accounting for energy prices, weather conditions, and occupancy predictions (Drgoňa et al., 2020). Despite its advantages, this approach has notable drawbacks related to computational complexity, dependence on accurate models, and costs for modelling, data management, expert monitoring, and deployment. MPC relies on precise mathematical models and substantial computational resources to predict and optimize system performance. Additionally, the success of MPC is contingent on the accuracy of these models, making their development and maintenance both time-consuming and costly (Afroz et al., 2018).

Model-Free Control for GEBs

Model-free techniques can learn directly from data and modify their best tactics over time, without the need for an explicit model of the system they are trying to control (Fliess and Join, 2013). ANN, fuzzy logic, deep RL, RL, and hybrid systems that integrate these techniques are important examples of these technologies. The distribution of each model-free technique from 2015 to 2023 is shown in Figure 33.

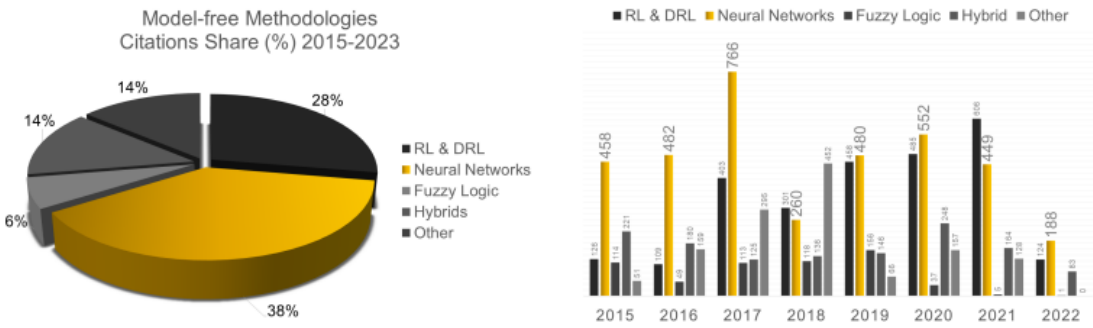

Figure 33. Model-free control citations share (%) and total citations count per methodology for HVAC control from the 2015-2023 period (Michailidis et al., 2024).

A specific area of model-free control, which is grounded in the mathematical structure of artificial neural networks (ANNs) such as RL, deep reinforcement learning (DRL) and deep artificial neural networks, holds promise for overcoming the challenges of energy management in complex and dynamic GEBs. This potential arises from the ability to continuously adapt and optimize control policies by learning from the actual outcomes of previous decisions, rather than relying on predicted models (Yu et al., 2020).

Moreover, model-free approaches tend to be more scalable, as they do not require explicit system models or continuous optimization. They can handle large, multi-variable systems with discrete or continuous nature without being overwhelmed by computational complexity, making them well-suited to complex GEBs (Pinto et al., 2021b).

While model-based approaches are less flexible and require frequent updates to handle new conditions, model-free approaches are more adaptable and improve automatically as new data becomes available through trial-and-error interactions with the control environment.
In general, model-free approaches offer greater flexibility, adaptability, and scalability, making them well-suited for complex, dynamic environments where traditional models are difficult to create. While they require time to learn and may initially be less accurate, model-free methods are capable of continuous improvement through data-driven learning, providing effective control in the long run. As GEBs become more integrated with variable RES, EVs, and smart grids, model-free approaches are potentially to play an increasingly important role in energy management systems (Yu et al., 2021a).